Author/s:
Bhatia, Shraey

Title:
Detection and Analysis of Climate Change Scepticism

Date:
2024-01

Persistent Link:
https://hdl.handle.net/11343/345314

Detection and Analysis of Climate Change Scepticism

Shraey Bhatia

ORCID: 0009-0003-6006-6028

School of Computing and Information Systems

University of Melbourne

This dissertation is submitted for the degree of

Doctor of Philosophy

May 2024

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 100,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Shraey Bhatia

May 2024

Preface

Large portions of Chapter 3 have appeared in:

> Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "Topic intrusion for automatic topic model evaluation." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 844-849. 2018.

Some portions of Chapter 4 have appeared in:

> Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "You are right. I am ALARMED–But by Climate Change Counter Movement." arXiv preprint arXiv:2004.14907 (2020).

Large portions of Chapter 5 have appeared in:

> Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "Automatic classification of neutralization techniques in the narrative of climate change scepticism." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2021.

Some portions of Chapter 6 have appeared in:

> Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "Automatic claim review for climate science via explanation generation." arXiv preprint arXiv:2107.14740 (2021).

Acknowledgements

Looking back on the past 5.5 years of my PhD journey, it's been quite a ride with its fair share of ups and downs, and some tough personal and health challenges. But, it's also been a time of incredible growth for which I'm really thankful. I couldn't have made it through without the help of some amazing people.

First and foremost, I extend my heartfelt gratitude to my supervisors, Prof. Timothy Baldwin and Dr. Jey Han Lau. Their unwavering motivation, expert guidance, and immense knowledge have been the cornerstone of my academic journey. They have stood by me as pillars of support, embodying the ideal mentors I could have ever wished for. During moments of doubt and uncertainty about my ability to complete this journey, their faith and patience were unyielding. Tim has been an inspirational figure, someone I have always looked up to with great admiration. Jey with his approachability and insightful mentorship, has been instrumental in fostering my intellectual growth, encouraging innovative ideas, and patiently clarifying even my most basic or seemingly trivial queries.

I am also thankful to my fellow research colleagues in the lab. Our stimulating conversations and discussions have been invaluable, contributing significantly to my academic and personal development.

A special note of gratitude goes to all the medical staff and doctors who provided me with exceptional care during some of the most challenging times, especially amidst the COVID pandemic and my battle with long-COVID. There were long periods of time when I was unable to work, clouded with uncertainty about returning to full health. However, their relentless support and care played a crucial role in my recovery, enabling me to complete my thesis.

Most importantly, my deepest appreciation goes to my family, who have been my unwavering support system. My father, whose motivation helped me push through the toughest times, and my mother, whose presence was a constant source of comfort during stressful periods. My gratitude also extends to my younger sister, whose lighthearted antics never failed to bring a smile to my face. Last but certainly not least, I must mention Ralph, my loving dog, who has been an incredible source of joy and a perfect antidote to stress.

This journey, with its highs and lows, has been an enriching experience, and I am profoundly thankful to everyone who has been a part of it. Your support and belief in me have been the driving forces behind this achievement. Thank you.

Abstract

Climate change, predominantly driven by human activities, poses a threat through effects like rising sea levels, melting ice caps, extreme droughts, and species extinction.The IPCC's 5th and 6th reports highlight the urgency of limiting global warming, with the latter projecting a concerning 1.5°C rise by 2040. Despite scientific consensus, the digital sphere is inundated with content that fuels scepticism, often sponsored by specific lobby groups. These articles, under the umbrella term of climate change scepticism (CCS), weave a blend of misinformation, propaganda, hoaxes and sensationalism, undermining collective climate action. This thesis aims to offer strategies to address this misleading narrative.

In this thesis we probe CCS through 4 dimensions: (1) understanding the underlying themes in the data, (2) detecting CCS articles, (3) understanding and detecting the framing and neutralization tactics used to construct CCS narratives and (4) fact-checking the veracity of claims, elucidating reasons for potential inaccuracies.

A notable challenge in addressing the aforementioned tasks is the limited availability of data. Throughout this thesis, we leverage advancements in natural language processing (NLP) to mitigate this. Pre-trained language models (PLMs) and their scaled counterparts, large language models (LLMs), have revolutionized our capacity to comprehend and generate text that mirrors human language. These models, adept at learning from real-world knowledge and semantics from extensive datasets prove extraordinarily effective over a diverse range of language tasks.

Topic models distil document collections into key themes, represented by groups of words or "topics" without the need of human labelling or any a priori notion of the content of collection. In essence, they offer a means of exposing underlying themes in the documents.

Each document typically aligns with one or several themes, but capturing the essence of the collection's context remains a challenge. In this thesis, we introduce methods that enhance the quality of topic outputs to better mirror the context of document collections.

For detection of CCS articles, there was no dataset available in this domain. We bridge this gap by scraping and compiling a dataset articles known to exhibit climate change scepticism. By extending training of PLMs on this dataset, we enhance their ability to discern stylistic and linguistic elements of CCS which allows the models to not only distinguish between CCS and non-CCS articles but also to highlight misleading spans indicative of scepticism.

To delve deeper into the intricacies of CCS narratives, we must analyze their argumentative framing. This is accomplished by employing techniques of framing and neutralization which translates it into a multi-task classification task. We propose an annotation task and collect human judgements. Given that data collection can be resource-intensive, we leverage unlabelled data in a semi-supervised setting achieving substantial performance gains.

Finally, we dive into the task of explanation generation to detail the reasons behind a claim's inaccuracies. Using LLMs in a retrieval-augmented approach, we connect the LLM to an external knowledge source like peer reviewed papers via a retriever. This retriever fetches pertinent "facts" related to the claim, enabling the LLM to both verify and explain the claim grounded to these facts. LLMs are prone to generate ungrounded information, commonly referred to as "hallucinations". We investigate approaches to detect such inaccuracies, then introduce methods to reduce these hallucinations, and finally employ LLM-based evaluations to assess the quality of the produced content.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Introduction

Climate change constitutes a critical global challenge, as evidenced by phenomena such as rising sea levels, melting polar ice caps, altering weather patterns, extreme droughts, and the extinction of countless species. The 5th assessment report by the Intergovernmental Panel on Climate Change (IPCC) unambiguously identifies human activities as the primary driver of this crisis, urging the containment of global warming to less than 2°C.[1] Moreover, the 6th report projects a likely 1.5°C increase before 2040, potentially escalating to a devastating 3°C rise by century's end.[2] This anthropogenic climate change has been a significant factor in events like the California fires (Goss et al., 2020), Australia's Black Summer bushfire disaster of 2019-20 (Abram et al., 2021; van Oldenborgh et al., 2020) and linked to other extreme weather patterns, including the 2021 floods in Germany and Belgium (Kreienkamp et al., 2021). Regrettably, it is not uncommon to see claims of questionable scientific merit with headlines such as *Climate Change has caused more rain, thereby aiding in fighting Australian wildfires.* Such narratives, lacking in scientific rigor, contribute to skepticism (Oreskes and Conway, 2011), propagate misinformation (Farrell, 2019), and neutralise and dilute crucial

[1] https://www.ipcc.ch/report/ar5/wg1/

[2] https://www.ipcc.ch/assessment-report/ar6/

... Prof. Dr. Ulrich Kutschera told in an interview that CO_2 is a blessing for mankind and that the claimed 97% consensus among scientists is a myth. ... he rejected extremes, among them the climate alarmists who predict a fictitious, imminent earth heat death ...
The CO_2 is not much of a concern and nothing for the oceans. Our harmless emissions of trifling quantities of CO_2 cannot acidify oceans
New Zealand schools to terrify children about the climate crisis. Who cares about education if you believe the world is ending? What will it take for sanity to return? Global cooling? Another Ice Age even? The climate lunatics ... encourage them to wag school to protest for more action.
Meanwhile, Australians are suffering from high energy prices now. Retail electricity prices have risen by more than 120 per cent in real terms over the past decade, while wholesale prices have tripled in the last three years. These price rises are primarily the result of heavy-handed government interference supporting renewables though the Renewable Energy Target at the expense of more reliable, affordable coal-fired power.

Table 1.1: Example of snippets from CCS articles

debates (McKie, 2018), thereby politicizing climate change (Benegal and Scruggs, 2018; Van der Linden et al., 2017) and hindering necessary action.

At the same point it is important to acknowledge that there is an increase in energy needs as countries like China and India continue to grow, and these are real needs which need to be met. Even proponents of sustainability can engage in unscientific thinking to advance their agendas. For instance, there is misleading information suggesting that solar energy output on cloudy days is almost equivalent to sunny conditions and that England could be entirely powered by solar energy; however, both are inaccurate due to reduced efficiency in cloudy conditions and the impractical land area required for such a solar setup in England (MacKay, 2016). Climate change and sustainability are ultimately complex issues, and everyone is susceptible to biases that can overshadow rational thought. In the remainder of this thesis, we will primarily focus on the debate surrounding climate scepticism, though we acknowledge that there are broader related issues such as energy needs that are also just as important in the discussion of climate change.

(a) Example 1

(b) Example 2

Figure 1.1: Articles exhibiting climate change scepticism from the Australian

This proliferation of misleading information is largely driven by counter-movement organizations, including fossil fuel lobbies, conservative think tanks, large corporations, and certain media outlets. These entities often challenge the established science of climate change, fueling climate change skepticism (CCS) articles (Boussalis and Coan, 2016; Dunlap and Jacques, 2013; Farrell, 2016; McKie, 2018; Oreskes and Conway, 2010). Digital media outlets, in particular, have played an instrumental role in this spread, as depicted in the 2 examples given in Figure 1.1. The narrative structure of CCS articles commonly merges misinformation, frivolity, propaganda, and hoaxes sprinkled with stylistic elements of sensationalism, melodrama, and clickbait. We present examples of these tactics in Table 1.1. The first example contains misinformation as it talks about CO2 being beneficial, thus promoting alternative facts, disregarding any of the scientific consensus. The second example trivializes the issue and uses lack of seriousness of the issue. Example 3 focusses on blame and utilizes sensationalism, a style of writing that features emotionally charged words, whereas in example 4, we see there is less focus on the stylistic components of sensationalism but more questioning of the economics and cost of renewable technologies? Public perception is shaped by the narratives in popular media and these narratives often gain momentum rapidly within social circles. These narratives disseminate information designed to captivate or retain attention, often without any validation of authenticity. We show how these narratives spread in Figure 1.2, where a news article wrongly suggesting that arsonists were responsible for the Australian fires gained traction and was further amplified by notable figures such as Donald Trump Jr., leading to widespread dissemination and ultimately undermining the role of climate change as the cause of the fires. This motivates the need for development of applications that can analyse these narratives and potentially inform readers of their dubious origins.

In this thesis, we breakdown the problem into four sub-problems, each addressing a specific aspect of the challenge. Firstly, we delve into understanding the underlying themes present in the dataset which involves identifying the most effective method to optimize the topic analysis of the corpus. Secondly, we focus on the detection of climate change skepticism articles which entails developing a mechanism to differentiate between CCS and non-CCS

Figure 1.2: Examples of scepticism spread

articles within a given collection. Thirdly, we explore the nuances embedded in CCS articles through the means of strategies such as framing and neutralization tactics, which are often

employed in these articles. Finally, the fourth aspect revolves around the fact-checking of claims made within these articles. Here, we aim to not only assess the veracity of these claims but also to elucidate the reasons behind potential inaccuracies, drawing support from trusted sources.

One common challenge in these tasks is the scarcity of data especially labelled data, driving the need for strategies like semi-supervised learning and transfer learning. Significant advancements in the field of Natural Language Processing (NLP), particularly regarding pre-trained language models (PLMs) and large language models (LLMs). Pre-trained models are sophisticated deep learning architectures, designed to capture real-world knowledge, semantics, and language structures from large datasets. These models are first trained on text data with a simple objective (e.g. to predict the next word given a sequence of words) and can then be adapted to tackle specific problems (e.g. categorise the topic of a document). Similarly, large language models — effectively a scaled up version of pre-trained models possess the capability to understand and generate text that closely resembles human language, thus proving invaluable for tasks such as summarization, question answering and information extraction. Furthermore, these models can be adapted to specific domains and connected to specific knowledge sources, thereby enhancing their utility in handling specialized information. Thus in this thesis, we aim to address the aforementioned questions, by applying NLP techniques to gain a more nuanced understanding of the narratives promoted by anti-climate change organisations.

Topic models has been a widely employed tool for analyzing discourse and themes within document collections. Topic models consist of topics which are typically represented by their top n words and individual documents are assigned a mixture of topics. Figure 1.3 gives a conceptual overview of topic models. A key attribute of an effective topic model is its ability to generate coherent topics, meaning that the set of n words should convey a clear and meaningful concept. Additionally, the quality of the topics produced should accurately reflect the context of the document collection. Achieving this involves hyperparameter optimization, a process historically guided by measures such as perplexity or topic coherence (Aletras and Stevenson, 2013; Lau et al., 2014; Mimno et al., 2011; Newman et al., 2010) which assess

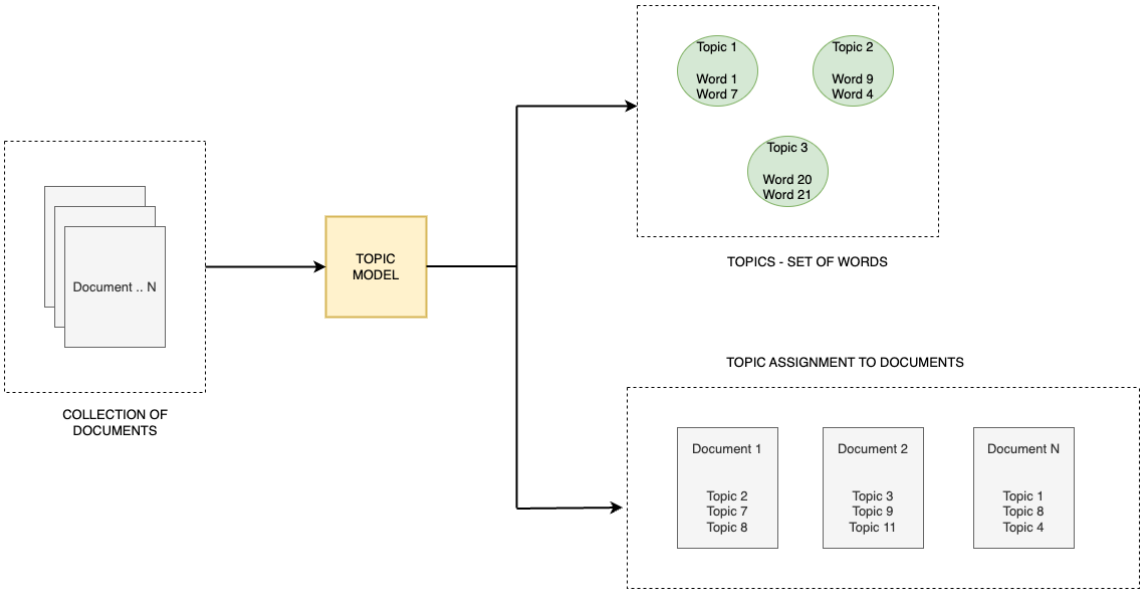


Figure 1.3: A conceptual overview of Topic Models

how well topic words relate to each other. However, relying solely on topic coherence for optimization may not provide a comprehensive understanding of the document collection. While it may result in highly coherent topics, these topics might offer limited collective insight into the broader themes and concepts present in the documents. Therefore, it is crucial for topic models to not only produce coherent topics but also to capture the essence of the themes and concepts within the document collection, thus necessitating the need for the development of strategies to achieve such a balance. In this thesis, we specifically focus on enhancing these strategies, especially considering the significant role of topic models in analyzing public discourse and opinions on climate change and tracking the evolution of climate-related themes over time (Boussalis and Coan, 2016; Dahal et al., 2019; Elgesem et al., 2019).

A fundamental aspect of NLP is understanding the nuances of language, for example to identify climate change skepticism. This capability to distinguish genuine information from scepticism is imperative. Pre-trained models, due to the extensive real-world and semantic knowledge embedded in their parameters, are potentially promising for detecting climate change scepticism but also miss the specific data distribution of this specialised

Document Text
Take a Look at the New Consensus on Global Warming. A scientific consensus has emerged among top mainstream climate scientists that "skeptics" or "lukewarmers" were not long ago derided for suggesting — there was a nearly two-decade long "hiatus" in global warming that climate models failed to accurately predict or replicate ... More importantly, the paper discusses the failure of climate models to predict or replicate the "slowdown" in early 21st century global temperatures ... Democrats and environmentalists praised Karl's work which came before the Obama administration unveiled its carbon dioxide regulations for power plants ... Then, in early 2016, mainstream scientists admitted the climate model trends did not match observations – a coup for scientists like Patrick Michaels and Chop Knappenberger who have been pointing out flaws in model predictions for years.
New Zealand schools to terrify children about the climate crisis. Who cares about education if you believe the world is ending? What will it take for sanity to return? Global cooling? Another Ice Age even? The climate lunatics ... encourage them to wag school to protest for more action.

Table 1.2: Climate scepticism spans in Red refer to spans with high scepticism ; Orange highlights scepticism but not as high as red, and black is non sceptic parts.

domain (Gururangan et al., 2020). Thus we can train it further over data specific to the domain to adapt it to the linguistic characteristics and semantic context of that domain, such as finance (Araci, 2019), medicine (Rasmy et al., 2021), or climate (Webersinke et al., 2021), thereby enhancing their performance within those specific areas - a process termed "domain adaptation". For instance, we could train it further on CCS articles to aid in to better understanding the language and narrative around CCS and use it to differentiate between CCS and non-CCS content. While identifying if an article is heavy on CCS is important, the subsequent challenge is to locate specific text spans manifesting skepticism. This scepticism can exist across a wide spectrum, from highly sceptical, emotionally charged statements to subtler, less dramatic instances, as illustrated by the examples in Table 1.2.

Detection of CCS articles provides a valuable initial step towards understanding of climate change scepticism, but to have a more nuanced understanding of CCS narratives, it's imperative to analyse how they frame the arguments. Framing is a method used in social sciences to group, organise and highlight aspects of an issue to make them salient (Entman,

Example	**Frame**
There's no indication this is anything but just natural variability, humans not playing a part	SCIENCE
Despite forecasts of warming the world has actually been cooling, so global warming is a hoax	SCIENCE
Renewable energy is way too expensive	POLICY
... New Zealand's actions should be less ambitious than Australia's because it is a wealthier country	POLICY

Table 1.3: Examples of counter climate arguments and their frames

1993). For instance in the context of CCS, framing can typically be categorized into two broad groups: Science and Policy. As the name suggests the 'Science' category deals with arguments based on the questioning of scientific facts and 'Policy' frame arguments target issues of cost and economy or pass the blame to other payers. We present a few examples of the science and policy frames in Table 1.3. Another method to understand narrative in social science is to analyse its "neutralisation" strategy, i.e. identify the justification for the deviant behaviour (Maruna and Copes, 2005; Sykes and Matza, 1957). A common example is the phrase *The cure can't be worse than the disease/problem* which has been used across a broad spectrum of issues including climate change skepticism and the COVID 19 pandemic.[3] In terms of climate change it could be thought that spending money on renewables is not worth it for the economy. Although originating in criminology, neutralization has been applied to various fields, such as corporate social responsibility (Fooks et al., 2013), fast fashion (Joy et al., 2012), the tobacco industry (Fooks et al., 2013; Oreskes and Conway, 2010), and CCS (McKie, 2018). Referring back to Figure 1.1, we observe applications of these ideas in CCS narratives. For instance, Figure 1.1a (*Opinion : China has zero ...*) passes the blame to other countries and suggests that they not us should be acting, based on the neutralization strategy popularly known as "justification by comparison". Similarly, looking at Figure 1.1b (*Climate Change has ...*) it denies the existence of climate change and global warming and

[3]https://www.business-standard.com/article/international/trump-opposes-perpetual-lockdown-says-cure-cannot-be-worse-than-problem-120101300184_1.html

CLAIM

Earth about to enter 30-YEAR 'Mini Ice Age'

VERDICT

INCORRECT

SOURCE: Harry Pettit, Sean Martin, Express, The Sun, 2 Feb. 2020

DETAILS

Factually Inaccurate: The most recent forecast from NOAA's Space Weather Prediction Center (from December 2019) predicts that the next solar cycle will be similar to the one that is currently ending.
Misleading: Even if an extended "grand solar minimum" were to occur, it would not produce marked global cooling.

KEY TAKE AWAY

Scientists cannot predict whether grand solar minimum, which is a decades-long period of lower solar activity, is coming. But even if one occurred, the consequences for average global temperatures would be minimal. Human-caused greenhouse gas emissions will continue to impact average temperatures much more strongly than solar activity cycles.

Figure 1.4: An example of a claim review from climatefeedback.org

accepts it as a fact of life, — a strategy popularly known as "denial of victim ". Traditionally, social scientists have manually analysed CCS articles and assigned the appropriate frame and neutralisation categories to the text spans in articles. But this is laborious and does not scale, and underscores the need for automated approaches and strategies for classifying CCS texts according to frames and neutralization strategy.

Going beyond identifying how CCS articles frame their arguments, the next natural step would be explaining why or how the claim is inaccurate. That is, our goal is to fact-check claims to verify their truthfulness and give a justification as to why they may not be truthful. Scientists and experts have have been doing this by manually supplying feedback for such bogus claims, verifying their truthfulness and offering the public scientifically sound information. Efforts to fulfill this mission have led to the publication of expert feedback on

websites like climatefeedback.org and skepticalscience.com. Figure 1.4 gives us one such example where the claim from The Sun *Earth is about to enter 30-year 'Mini Ice Age'*, has been labelled as Incorrect, with the "Key Take Away" being that *Scientists cannot predict whether solar grand minimum ... is coming* and *even if one occurred, the consequences for average global temperatures would be minimal*. This multifaceted process of claim verification, coupled with textual explanation and justification, is one we aim to automate. Such a tool would enable climate science experts to respond more efficiently and on a broader scale to more claims. Our idea is to connect LLMs to trusted sources and build a system that can verify and explain the truthfulness of a claim based on these sources. In other words, given a claim the system would find relevant facts from authoritative sources and then create an accurate and clear explanation based on those facts.

1.2 Research Questions and Contributions

The research questions addressed in this thesis, and its contributions, are as follows:

Topic Models:

- Can we automate the evaluation of topic models based on how well they allocate topics to documents in the collection?

The evaluation of topic models is generally addressed through both intrinsic and extrinsic means, examining perplexity for intrinsic evaluation and topic coherence for extrinsic evaluation. We propose an alternative approach to topic model evaluation, by assessing whether the topics assigned to the documents are meaningful through document-level evaluation of topic intrusion, based on the setup first introduced by Chang et al. (2009). In topic intrusion, users are presented with a document, a set of allocated topics by a topic model and an intruder topic, and they must identify the intruder. The intuition is that if the user fails to recognise the intruder topic, then the allocated topics do not capture the document. Recognizing that collecting these human judgements can be both time-consuming and expensive, we introduce an approach employing a neural network model with additional corpus level features (e.g.

frequency of words in the corpus) to automate the task. Finally, we propose a new metric to measure the performance of the approach and use this metric to rank topics produced by topic models.

CCS detection

- How can we use pre-trained language models to detect CCS articles?
- Can we extend our approach to further to highlight the most misleading text spans in CCS articles ?

The initial step in our modelling process requires the development of a suitable dataset i.e. a collection of articles that are known to exhibit CCS. At the time we embarked on this research, there was no such dataset specifically tailored for Climate Change Scepticism. To address this gap, we scraped articles published by various climate change counter-movement organizations. To ensure the robustness of our model, we also meticulously curated a test set from a variety of CCS and non CCS sources encompassing diverse styles, such as news, sensationalism, and satire. Based on this dataset, we frame the task as a binary or a 1-class classification problem i.e. given an article, classify whether the article exhibits CCS or not. Given that a (trained) language model produces a probability estimate for a sequence of words (which measures how much the text fits the training data), we explore adapting a PLM by fine-tuning it on CCS articles. Post adaptation, we can decide whether an article is CCS or not by looking at its probability estimate for the article (high probability indicates an CCS article). Envisioning its application as an alert system for flagging climate change skepticism in web articles, we extend our approach to highlight specific content that espouses these skeptical views to users. The advantage of using a fine-tuned PLM for the detection task is that we can examine individual word probabilities, enabling us to detect "spans" of text that look most like CCS text.

Framing and Neutralization

- How can we automate the identification of frames and neutralization tactics used in the arguments of CCS articles?
- What is the impact of the quantity of labeled training data on classification performance, and can we include unlabelled data, that is semi-supervised learning to improve performance?

Next, we delve into the identification of more nuanced classes of arguments employed in the creation of CCS texts, achieved through the application of framing and neutralization techniques. We first introduce the task of neutralization as a multilabel classification task. Similar to the previous research question, the first step here revolves around data collection but unlike the previous task, neutralization and framing comprise multiple classes (2 for framing, further subdivided into 7 for neutralization) necessitating human annotation. To this end, we propose an annotation task for it and gather human judgements. However, data collection can be expensive and time-consuming, leading us to leverage unlabelled data in a semi-supervised multi-task learning objective. We show significant performance gains in this setting with the best models performing on par with human-level performance, and further demonstrate how this approach can be utilized to analyse the detected spans from the previous research question.

Claim Verification and Explanation generation

- How can we incorporate knowledge from trusted sources to automatically classify and explain the truthfulness of climate change claims?
- How can we manage and mitigate the occurrence of hallucinations — the phenomenon where new information not grounded in the trusted sources is produced in the generated explanations?

We introduce the idea of explanation generation as a means to justify the truthfulness label predicted by a fact-checking model. Here we explore the use of large language models,

as they are capable of comprehending and generating fluent text. To this end, we use retrieval-augmented generation approach, where the LLM is connected to an external knowledge source through a "retriever". This retriever queries the knowledge source to find relevant "facts" pertaining to a given claim and this information is provided to the LLM, so that it can verify and explain the claim grounded to the provided facts. Given that LLMs are prone to generating hallucinations, i.e. novel information that is not grounded in the presented facts, we delve into the metrics used to first identify hallucinations, then propose methodologies to mitigate hallucination by exploring the use of multiple knowledge sources and strategies such as paraphrasing. Finally, we use LLM-based evaluation to assess the quality of the generated outputs

1.3 Limitations

This thesis, conducted over a period of 5.5 years, coincides with substantial changes in the NLP landscape, particularly in the recent years with the emergence of LLMs such as ChatGPT,[4] GPT-4,[5] Claude, [6] and LLaMA (Touvron et al., 2023) to name a few. As such, we see an evolution of 'state-of-the-art' approaches/methods from chapter to chapter, and it's important to understand the context and time over which these studies were done. Although Chapter 2 briefly presents literature related to these developments, much of the recent research has not been employed in our work, as it postdates the work in the thesis.

For example, Chapter 3 focuses on topic models, and it uses a convolutional neural network as the backbone model, as PLMs had not been introduced at that time. In Chapter 4 and 5, PLMs such as GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019) are utilized, but not their more advanced versions like GPT-3/4 (Brown et al., 2020). In Chapter 6, more recent models like Flan–T5 (Chung et al., 2022) and GPT-4 are employed, along with approaches such as instruction tuning. In practice these LLMs could be applied to

[4]https://openai.com/blog/chatgpt
[5]https://openai.com/research/gpt-4
[6]https://www.anthropic.com/index/introducing-claude

previous chapters and might lead to performance gains; however, such experimentation was not undertaken in this thesis due to the timing of work in the respective chapters.

1.4 Thesis Structure

As this work situates itself in the computational social sciences, Chapter 2 is made up of 2 different sections. The first section focuses on the literature surrounding the narrative of climate change from a social sciences perspective. We begin with an introduction to the history of the climate change debate, a summary of scientific reports, and an exploration of the polarization and politicization of climate change. The chapter also delves into corpus linguistics, examining the topical and stylistic facets of climate change language, followed by a synopsis on Climate Change Counter Movements and their role in fostering skepticism. We further discuss theories around framing and neutralization that are often employed to fuel this skepticism. The second section centers on the approaches and methods used to automatically analyse or detect CCS, from the natural language processing (NLP) perspective. Here, we provide an overview of topic models, including their structure, types, and applications. We also examine PLMs, covering different training objectives, domain adaptation approaches, and fine-tuning strategies for specific domain data. We conclude the chapter with an exploration of the literature on their misinformation and propaganda detection, both in terms of novel datasets and methodologies and their relation to climate change.

In Chapter 3, we direct our focus towards topic models, and introduce an alternative to evaluating these models through the task of topic intrusion. We begin the chapter with a concise review of the literature pertaining to the evaluation of topic models. Subsequently, we explain the task of topic intrusion, along with the corresponding process for annotation collection. This is followed by the presentation of the adopted methodology, which is based on the utilization of a convolutional neural network enhanced with corpus level features. We conclude the chapter with the proposal of a novel metric designed to measure system performance.

In Chapter 4, we introduce the task of detecting CCS within documents, and develop a dataset consisting of CCS articles together with articles from various non-CCS sources. We then propose a novel system for detecting CCS articles, leveraging domain-adapted PLMs. This is followed by a comparison unidirectional and bidirectional models. Finally, we demonstrate that the detection methods can be modified to identify spans of potentially misleading texts, and end with a discussion on extending the methodology to accommodate short texts.

In Chapter 5, we delve into the theory of neutralization and framing, analyzing it through the lens of semi-supervised learning. We begin the chapter with a review of semi-supervised learning, encompassing its earliest methods, applications in social sciences, and the contemporary fusion of pre-trained models with semi-supervised learning approaches. Following this, we present framing and neutralization techniques for analyzing the narrative of climate change skepticism, e.g., whether it is justifying inaction or promoting alternative views. We introduce neutralization as a multilabel classification task, followed by the development of a dataset with manual annotations of NT. Finally, we explore a semi-supervised model that uses unlabeled data for the classification task, comparing its results with human performance, and highlighting the advantages of semi-supervised learning.

In Chapter 6, we delve into the task of claim verification, focusing on both veracity prediction and explanation generation. We begin the chapter with a review of retrieval-augmented generation, spotlighting early approaches in open-domain question answering and their applicability to our problem. We then provide an overview of the literature on the problem of hallucination in LLMs, examining the underlying causes and continue by summarizing the evaluation techniques used for generations and how they sync to our task. After that, we introduce the datasets of interest to our task such as Climate FEVER and external knowledge sources like Wikipedia. Subsequently, we detail our retrieval-augmented generation system, also including evaluation metrics like BERT-score. We also investigate hallucination in the generated explanations and explore strategies to mitigate them. Next, we demonstrate how LLMs can be used to measure the quality of the generated explanation, thereby unifying the discussion around claim verification, hallucination, and evaluation.

Finally, we test our best models with alternative knowledge sources like IPCC reports, and peer-reviewed articles.

We summarise the contributions of the thesis in Chapter 7 and present avenues for future work.

Chapter 2

Background

In this chapter, we present the background and literature around climate change, and broadly divided into 2 parts the social sciences and natural language processing (NLP). The first part delves into the social science aspect of climate change. We start by providing an introduction to the history of the climate change debate. This involves examining the study of climate patterns and the formulation of the problem of global warming, including the various assessment reports that have shaped our understanding of these issues over time. We then pivot to the movement of climate change skepticism, detailing its origin, unique characteristics of its communication strategies and the language employed. Subsequently we present the framing and neutralization literature used to construct the narrative around climate change scepticism which has played a key role in the public discourse on this subject. In the second part, we transition to the realm of NLP, which provides useful tools and perspectives for analysing the language and narratives around climate change. We commence our discussion with an overview of different topic models. We then delve into pre-trained language models, exploring various architectures and models, and discussing methodologies like fine-tuning and domain adaptation, which allow these models to be tailored for specific tasks or contexts. Finally, the chapter reviews the literature on misinformation and fact-checking. We examine the different datasets and techniques employed in this field and discuss how they sync with the domain of climate change.

2.1 Introduction and History on Climate Science

Climate change is one of the biggest challenges threatening the world, and we are at a defining moment. Rising sea levels, melting polar ice, changing weather patterns, severe droughts (Breshears et al., 2005), and extinction of species are just some of the dreadful effects of this crisis (Breshears et al., 2005; Gardner et al., 2018; Harrison et al., 2018; Lewis, 2006). The Intergovernmental Panel on Climate Change (IPCC) in its 5th assessment report categorically concluded that humans are the main culprit and there is a need to limit global warming to less than 2 C.[1] To achieve a level of consensus, review national communications, and establish emission targets, the Conference of the Parties (COP) was created. This decision-making body, part of an international convention, meets annually to review, negotiate, and coordinate measures to tackle climate change. COP has been held since 1995, with varying degrees of success and key milestones being COP3 in Kyoto, COP11 in Montreal, COP17 in Durban, and most notably, COP21 in Paris where agreement was reached to limit global temperature rise to well below 2°C (Rhodes, 2016). Despite these efforts, the 2022 Intergovernmental Panel on Climate Change sixth report warned of the necessity for drastic action to meet the goals set at COP21 and to keep global warming under 2°C.[2]

The history of studying climate patterns and phenomena can be traced back to ancient Greece (4th century BCE to 2nd Century BCE) and the time of Aristotle, Theophrastus and astronomers Eratosthenes and Ptolemy (Moser, 2010; Weart, 2010). In the 19th century, most of the debate and studies were around factors influencing local climate i.e. studies around land use for more or less rain, deforestation etc, rather than on global and human influence on climate patterns.[3] Later in the century, Tyndall (1872) explored radiation absorption especially with respect to water vapor and hydrocarbons like methane and carbon dioxide; this trapped radiation later, leading to the coining of the term “greenhouse effect". Arrhenius (1896) [4] tried quantifying CO2 emissions and calculated the estimated contribution

[1] https://www.ipcc.ch/report/ar5/wg1/

[2] https://www.ipcc.ch/assessment-report/ar6/

[3] See (Stehr et al., 1995) and (Fleming, 2005) that summarizes these 19th century studies

[4] The work was published again later in 2011 (Arrhenius, 2011)

of humans to CO2 levels in atmosphere, but the issue remained largely dormant and there was speculation that warmer temperatures could potentially be beneficial (Ekholm, 1901).

Moving forward to the 20th century, Callendar (1949) revisited the work of Arrhenius and provided evidence for the rise in CO2 levels in the atmosphere and argued this rise in CO2 was the cause of increased warming. Later, in a similar vein while studying seawater, Revelle and Suess (1957) ascertained that the oceans had a finite capacity to absorb CO2 (which countered the widely held belief that oceans can absorb all/infinite amounts of CO2) and hypothesised that there would be an increase of CO2 in atmosphere, which was later demonstrated by Keeling (1960) by taking snapshots of CO2 measurements 2 years apart. Sawyer (1972) summarised the science so far around anthropogenic global warming, and also predicted the rate of global warming for the next 3 decades, but around the same time the debate around global warming was overshadowed by the alternative debate around global cooling and the natural cycle of the ice age (Bryson, 1974; Kukla and Kočí, 1972). Manabe and Bryan (1969); Manabe and Wetherald (1967) gathered more evidence by developing a global climate model, and calculated that doubling CO2 would lead to a 2°C rise in global temperature. In the 1980s, consensus started building in the scientific community with Hansen (1988) giving one of the first assessments which suggested that humans had already affected the global climate. Eventually all these events led to establishment of Intergovernmental Panel on Climate Change by the World Meteorological Organization to provide regular scientific assessments on the implications and future risks of climate change which so far has presented 6 assessment reports (in 1990, 1995, 2001, 2007, 2014 and 2022).[5]

2.2 Climate Change Scepticism Movement

Anthropogenic climate change has been at the heart of the fires in California (Goss et al., 2020) and Black Summer bushfire disaster of 2019-20 in Australia (Abram et al., 2021; van Oldenborgh et al., 2020), which led to the destruction of 17 million hectares of land

[5]Parts of history on climate science were compiled later in the form or review papers, books and papers tracing historical works. See Fleming (2013); Keeling (1998); Nicholls (2007); Sherwood (2011)

... Prof. Dr. Ulrich Kutschera told in an interview that CO_2 is a blessing for mankind and that the claimed 97% consensus among scientists is a myth. ... he rejected extremes, among them the climate alarmists who predict a fictitious, imminent earth heat death ...
New Zealand schools to terrify children about the climate crisis. Who cares about education if you believe the world is ending? What will it take for sanity to return? Global cooling? Another Ice Age even? The climate lunatics ... encourage them to wag school to protest for more action.
Meanwhile, Australians are suffering from high energy prices now. Retail electricity prices have risen by more than 120 per cent in real terms over the past decade, while wholesale prices have tripled in the last three years. These price rises are primarily the result of heavy-handed government interference supporting renewables though the Renewable Energy Target at the expense of more reliable, affordable coal-fired power.

Table 2.1: Example of snippets from CCS documents reproduced from Table 1.1

and the death of a billion animals.[6] Jones et al. (2020) in their assessment concluded that human induced climate change is only going to increase the frequency and intensity of such fires, with Southeast Australia being highly vulnerable to these fire events (Dowdy et al., 2019). These climate change related disasters were not just restricted to fires in Australia and California but also to other extreme weather patterns like floods in Germany and Belgium in 2021 (Kreienkamp et al., 2021). During these times, we see articles with headlines such as “Climate Change has caused more rain, helping fight Australian wildfires" spreading misinformation to influence the narrative of climate change.[7] Oreskes and Conway (2011) argued that this narrative around climate change has inadvertently cultivated skepticism, resulting in a widespread lack of awareness and understanding of the issue. Ranney and Clark (2016) echoed the sentiment, suggesting this illiteracy exacerbates the challenges associated with effectively communicating the realities of climate change to the public (Moser, 2010). Anderegg et al. (2010) found that climate denial and the spread of misinformation have

[6] https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp1920/Quick_Guides/AustralianBushfires

[7] https://www.heartland.org/news-opinion/news/climate-change-has-caused-more-rain-helping-fight-australian-wildfires

a detrimental impact, ultimately leading to the discredit of climate science and scientists, which further jeopardises and influences how scientists engage with the public.

Despite the findings of the IPCC's 5th and 6th Assessment Report and a more than 97 percent consensus in the scientific community supports anthropogenic global warming (Cook et al., 2013), coordinated efforts to tackle the climate crisis are lacking. This can be attributed to the rise in opposing voices including the fossil fuel lobby, conservative thinktanks, big corporations, and digital/print media questioning the science and research around climate change. These climate sceptic organisations are collectively referred to as climate change counter movement organisations (Boussalis and Coan, 2016; Dunlap and Jacques, 2013; Farrell, 2016; McKie, 2018; Oreskes and Conway, 2010) and have been a major contributor to the proliferation of climate change scepticism articles. We collectively refer to them as CCS organizations and CCS articles. McKie (2018) argued that the motivation behind these CCS organisations is to maintain the status quo of the hegemony of fossil fuel-based neo-liberal global capitalism. These organisations are found around the globe and can masquerade as philanthropic organisations to fund climate misinformation (Farrell, 2019), hide behind libertarian ideas (McKie, 2018) to question scientists, and augment scepticism to promote pseudo science or "alternative facts". Some of these organisations have catchy names such as carbonsense.com or friendsofscience.org, and organise their own "scientific" conferences.

It is important that we understand the language and communication around climate change, as the way the public perceives and reacts to the constant supply of information around climate change is a function of how the facts and narrative are presented to them (Fløttum, 2014; Fløttum et al., 2016). Flottum (2017) emphasises that language and communication around climate change are significant, as climate is not just the physical science but has political, social, and ethical aspects, and involves various stakeholders, interests, and voices. A range of corpus linguistic methods have been used to study the topical and stylistic aspects of language around climate change. One such popular method is topic modelling which we discuss separately in detail in Section 2.4. Salway et al. (2014) leveraged unsupervised grammar induction and pattern extraction methods to find common phrases in climate change communication. Atanasova and Koteyko (2017) analysed frequently-used

metaphors manually in editorials and op-eds, and concluded that the communication in The Guardian (U.K.) was predominantly war based (e.g. *threat of climate change*), Seuddeutsche (Germany) based on illness (e.g. *earth has fever*), and the NYTimes (U.S.A) based on the idea of a journey (e.g. *many small steps in the right direction*).

In early works to study CCS and Conservative Think Tanks claims, McCright and Dunlap (2000) collected a sample of 224 documents from 14 different CCS and manually coded them to analyse the themes and arguments used in these contrary claims. Similarly, Elsasser and Dunlap (2013) sampled 203 op-eds from the US over a period of 2007- 2010[8] and broadly categorise them as being policy relevant vs non-policy issues, by employing an array of arguments assembled from *skpeticalscience.com*[9] (Washington, 2013). Sharman (2014) explored climate scepticism with the help of social network analysis, using node betweenness and degree of centrality to find and highlight central blogs and themes in the blogosphere. Boussalis and Coan (2016) proposed the use of probabilistic topic modelling based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) over a collection of 16000 documents from 19 CCS organizations collected over 15 years (1998-2013). The resulting topics covered themes ranging from sea and land impacts, scientific integrity, energy polices and politics around carbon trading, which were further coded by the authors to separate the science vs policy frames (in Section 2.3). More recently, Varini et al. (2020) released a dataset to study sentence based climate topic detection and demonstrated the advantage of neural network models over popular keywords based models.

Oreskes and Conway (2010) concluded in their analysis that the strategies employed by CCS organizations to construct the narrative to spread misinformation resemble the ones historically used by the tobacco lobby. For instance, targeting researchers and questioning the methodology of their research, and blaming scientific standards are strategies used by both CCS and tobacco lobby groups (McKie, 2018; Oreskes and Conway, 2010).

Several snippets from CCS articles are presented in Table 2.1. The first example, contains misinformation as it talks about CO2 being beneficial thus promoting alternative facts,

[8]Spanning 2 different administrations, the Bush administration followed by the Obama administration meaning there was a shift in polices and debate

[9]www.skepticalscience.com

Argument	**Frame**
CO2 is plant food and is good for the planet	Science
Climate change is natural and has always been changing	Science
We are entering another ice age	Science
Adapting to global warming is cheaper than preventing it	Policy
Renewable energy is way too expensive	Policy

Table 2.2: Examples of counter climate arguments and their frames.

disregarding any of the scientific consensus and its effects on climate change. The second example focusses on blame and utilizes sensationalism, a style of writing that features emotionally charged words. Looking at example 3, we can see the theme is around questioning the economics and cost of renewable technologies. Thus we hypothesise that CCS articles can be studied according to 2 devices: topical and stylistic. The topical aspects describe common issues discussed in CCS articles (e.g. carbon tax, fossil fuel, and renewable energy); the stylistic aspects capture how the narrative is presented — e.g. the use of exaggeration and sensationalism — similar to propaganda materials. There is no single strategy in CCS documents, but rather techniques in misinformation, propaganda, and neutralisation all play a role. Social scientists have explored various strategies to analyse these CCS narratives; one approach is the use of framing, a technique we will delve into in the following section.

2.3 Framing and Neutralization

Framing in social sciences is the process of grouping, organising and highlighting aspects of an issue to make them salient, thereby helping to define the problem more concretely, identify causes, inject moral evaluation and put forward possible solutions (Entman, 1993). But even earlier, Goffman (1974) first introduced the framing theory where he explained frames as "schemata of interpretation" that allow individuals to locate, perceive, and label events within their life space and the world at large. Baker et al. (1998) used framing in the context of NLP where they constructed a lexical database that associates words with semantic frames aligned to the work of Goffman (1974) serving as a key piece of political communication,

which has been applied to influence public opinion and thus having implications on policy framework and decisions and building on it (Chong and Druckman, 2007; Iyengar, 1994). Framing topologies and coding schema are generally divided into 2 different kinds of topologies, namely issue generic (Boydstun et al., 2013), issue specific (Benson, 2013) or some combination of both (Mendelsohn et al., 2021). Issue generic framing is not confined to a particular issue, but instead addresses broader themes or viewpoints that can be used across different subjects. An example, is the issue-generic framing of "economic consequences", which could apply to a wide range of topics like immigration policy, healthcare reform or environmental regulations. In contrast, issue specific framing refers to framing topologies that are specific to a particular issue. For instance, a specific framing for the issue of climate change might include topics like "rising sea levels," "carbon footprint," or "renewable energy". These are all topics that are specifically related to the larger issue of climate change and would not necessarily be applicable to other issues. These topologies been used extensively in mass media, social media (Card et al., 2016; Field et al., 2018; Kwak et al., 2020) or more recently to study the discourse around immigration (Mendelsohn et al., 2021). Boussalis and Coan (2016); Dunlap and Brulle (2015); Farrell (2016) categorised CCS arguments into 2 frames: "science" and "policy". Science frame arguments question the scientific facts, and deliberately plant doubt to sway the public towards pseudo science, whereas policy frame arguments target issues of cost and economy (e.g. carbon tax) or pass the blame for action to other nations. We present several examples of arguments in the science and policy frames in Table 2.2

Framing gives us the broader grouping of the narrative but to dive deeper into the strategies used in the narrative we need to understand the theory of neutralization. *The cure can't be worse than the disease/problem* is a phrase frequently used by climate change sceptics,[10] and also recently by Donald Trump in reference to COVID-19.[11] Though two widely different issues, *neutralization* is used to justify opposing a policy, lack of action, and thus promotion of either total denial of the problem (Diethelm and McKee, 2009) or its

[10] https://www.wired.com/story/the-analogy-between-covid-19-and-climate-change-is-eerily-precise/

[11] https://www.business-standard.com/article/international/trump-opposes-perpetual-lockdown-says-cure-cannot-be-worse-than-problem-120101300184_1.html

severity. In social science, neutralization is defined as justification/vindication for a deviant behaviour (Kaptein and Van Helvoort, 2019; Maruna and Copes, 2005; Sykes and Matza, 1957). Though initially developed in the field of criminology, it has been widely extended to other fields. For example, Fooks et al. (2013) studied it in through the lens of lack of corporate social responsibility. Similarly, Delmas and Burbano (2011); Lynch et al. (2010); Lynch and Stretesky (2013) explored elements of it in green washing, corporate greening and fast fashion (Joy et al., 2012), where consumers are misled over environmental performance, or justifications are provided for a poor environmental record. In a similar vein, Fooks et al. (2013); Oreskes and Conway (2010) studied neutralization in the tobacco industry, where campaigns were launched to blame and spread scepticism around scientific standards based on research on the public health concerns. McKie (2018) extended it to the domain of climate change. Table 2.3 presents examples of neutralization in the domain of climate change. In the first example the justification for lack of action is provided by a false equivalence that the policies could be counter productive for the poor, thus using the phrase "ridding the patient of the disease but only by killing them." In the second and third example the justification is provided by blaming "alarmist greens" and "IPCC" respectively and at the same time arguing about global warming being a natural cycle implying that it has nothing to do with humans.

To study neutralization on a fine grained level this general concept needs more sub-categorization. Sykes and Matza (1957) first introduced the techniques of neutralization, known as the "famous five", namely (1) Denial of Responsibility; (2) Denial of Injury or Harm; (3) Denial of Victim; (4) Condemnation of Condemner; and (5) Appeal to Higher Loyalties as tools for justification of deviant behaviour. The neutralization technique inventory has since been expanded to include "metaphor of ledger" (Klockars, 1974), that current deviant behaviour should also take into account past or future good behaviour , "dispersal of blame" (Thompson, 1980), where blame and responsibility is divided among the group rather solely on the shoulders of an individual; "defence or necessity or excuse acceptance" (Minor, 1981) where absolute necessity of the actions is falsely asserted and with no other choice left; and "no one cares" (Shigihara, 2013). More recently, Kaptein and Van Helvoort

Sure, we should reduce greenhouse gases, but if our climate policies hurt our ability to create more wealth and bring power to the world's poor, then we are ridding the patient of the disease, but only by killing him
It's very convenient for alarmist greens to blame the fires of Australia and California on global warming. In reality, global warming is just a natural cycle and the policies they themselves advocate are the culprits.
The IPCC falsely attributes natural warming and urban warming to greenhouse gas (GHG) emission warming. It ignores the compelling evidence of natural climate change before 1950 that correlates well with indicators of solar activity

Table 2.3: Neutralization examples

(2019) developed a hierarchical schema which combines these strategies into categorizations and sub-categorizations.

As mentioned earlier, Dunlap and Brulle (2015), Farrell (2016), and Boussalis and Coan (2016) categorised CCS arguments into 2 frames: science ("science") and policy ("policy"). McKie (2018) adapted Sykes and Matza (1957)'s original neutralization categories to analyse CCS narrative and connected them with the 'Science' and 'Policy' frames as follows:

- **Denial of Responsibility (Deny-Responsibility $\rightsquigarrow$Science)**: climate change is happening, but is a natural cycle and humans are not responsible.
- **Denial of Injury1 (Deny-Injury1 $\rightsquigarrow$Science)**: there are no significant harms attributable to climate change, and claims are generally overstated.
- **Denial of Injury2 (Deny-Injury2 $\rightsquigarrow$Science)**: there are benefits in rising C02 levels which have a positive effect on the environment.
- **Denial of Victim (Deny-Victim $\rightsquigarrow$Science)**: there is no evidence of climate change and no climate change victims; total denial of any global warming.
- **Condemnation of the Condemner (Condemn $\rightsquigarrow$Policy)**: climate change is misrepresented by scientists or manipulated by politicians, the media, environmentalists, etc.

Argument or Example	**NT**	**Frame**
There's no indication this is anything but just natural variability, humans not playing a part	Deny-Responsibility	Science
there is the very real probability that the global warming been overestimated by computer models, so warming is not too bad	Deny-Injury1	Science
CO2 is plant food and is good for the planet, also essential for plants in photosynthesis	Deny-Injury2	Science
despite forecasts of warming, the world has actually been cooling, so global warming is a hoax	Deny-Victim	Science
an avalanche of global warming alarmism is about to hit, thanks to environmentalists, media and a few scientists	Condemn	Policy
So-called "new renewable energy technologies" are extremely expensive and rely on huge subsidies, rising energy costs	Loyalties	Policy
... New Zealand's actions should be less ambitious than Australia's because it is a wealthier country	Justify	Policy

Table 2.4: Examples of counter climate arguments and their frames.

- **Appeal to Higher Loyalties (Loyalties $\rightsquigarrow$Policy)**: economic progress and development are more important than action on climate change, and hence policies like renewables or carbon taxes are not worth it.
- **Justification by Comparison (Justify $\rightsquigarrow$Policy)**: our actions are not as important as other countries which pollute more, or there are other more important issues than global warming.

We give examples of these 7 neutralization techniques in Table 2.4. Going back to Table 2.3 we can now analyse the examples through the lens of these neutralization techniques. CCS texts often use multiple neutralization techniques together. For example, the second example in Table 2.3 uses 'Condemn' (Policy) to blame the *alarmist greens* and 'Denial of Responsibility' (Science) to highlight that global warming is a *natural cycle* and the third example uses 'Condemn' (Policy) to accuse the IPCC of false attribution and 'Denial of Responsibility' (Science) to point out climate change being natural and linked to solar activity.

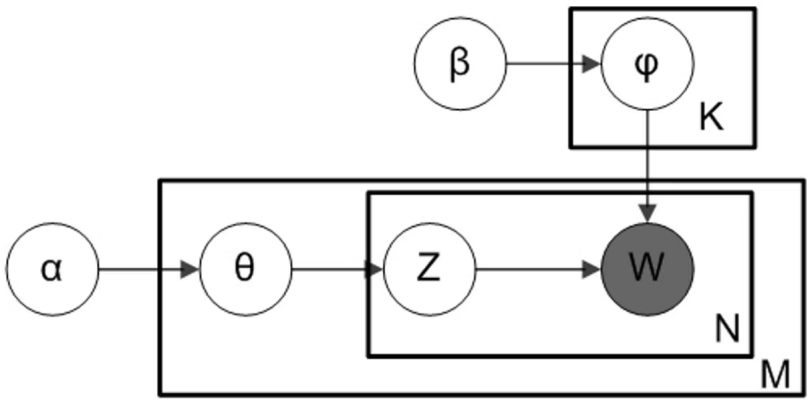


Figure 2.1: Plate Notation of LDA

Next, we will delve into various methodologies and architectures in NLP, to discuss how they are used to analyse and understand climate change narrative.

2.4 Topic Models

Topic modelling is a crucial tool in NLP that assists in unearthing hidden thematic structures within vast volumes of unstructured textual data. Topic models provide an efficient way to extract a collection's latent themes and concepts, referred to as "topics", making them invaluable in a wide array of applications ranging from document classification to information retrieval. Topic models jointly learn latent topics as a multinomial distribution over words and topic distributions for each document in the collection in the form of multinomial distribution over words.

Deerwester et al. (1990) introduced Latent Semantic Analysis (LSA), one of the earliest topic models which operates by constructing a term-document matrix — a matrix of the size of unique terms by the number of documents, traditionally weighted using tf-idf. To handle the inherent sparsity of the term-document matrix, LSA employs Singular Value Decomposition (SVD), enabling a reduction in dimensionality that still captures the salient features of the document collection.

Building on LSA, Blei et al. (2003) proposed Latent Dirichlet Allocation (LDA), one of the most widely used topic models. LDA involves the use of Dirichlet priors (α and β)

which serve as conjugate priors for the word-topic (ϕ) and topic-document (θ) multinomial distributions. A visual plate notation of LDA is given in Figure 2.1. The variables are:

- K is the number of topics
- N is the total number of words in the document collection
- M is the total number of documents
- ϕ is the word-topic distribution
- θ is the topic-document distribution
- Z is the topic assignment for the word
- W is the word in document
- α is the Dirichlet prior for topic distribution in documents
- β is the Dirichlet prior for word distribution in topics

LDA follows a generative process that consists of drawing $\theta(d)$ for each document and $\phi(z)$ for each topic from their respective Dirichlet distributions, then drawing a topic z from $\theta(d)$, and finally drawing a word from $\phi(z)$. However, exact inference i.e. calculation of posterior in LDA, is intractable due to the coupling between the latent variables in the joint distribution, which motivates the use of approximate inference methods, such as Collapsed Gibbs Sampling (Griffiths and Steyvers, 2004) and Variational Inference (Blei et al., 2017; Teh et al., 2006). Gibbs Sampling is from the school of Markov Chain Monte Carlo (MCMC) methods and operates by iteratively sampling from the conditional distribution of each latent variable, given all the other latent variables, and asymptotically converges to the true posterior distribution. One of the shortcomings of Gibbs Sampling is that it requires many iterations to converge, especially for large datasets or complex models. On the other hand, Variational Inference treats calculation of posterior as an optimization problem and aims to find an approximation to the true posterior that minimizes the Kullback-Leibler divergence (Blei et al., 2003) making the approach deterministic and generally faster than Gibbs Sampling.

But the limitation of variational inference is the approximated posterior can sometimes be substantially different to the true posterior, especially in scenarios when the true posterior is multimodal or highly skewed (Wainwright et al., 2008).

One of the limitations of LDA is that it assumes topic independence, which was addressed by Correlated Topic Model (CTM) an extension of LDA proposed by Blei and Lafferty (2006a). CTM models topic correlations and reduces topic content overlap by employing a logistic normal prior over topic proportions, in place of a Dirichlet prior. This shift, influenced by a multivariate Gaussian distribution, induces dependencies between topics. LDA being a parametric model however, lacks hierarchy and requires predefined topic numbers. Buntine and Mishra (2014) presented a tool kit HCA which introduced model burstiness and distinct priors on document-topic and word-topic distributions, unlike symmetric Dirichlet priors in LDA. Symmetric priors enable efficient collapsed Gibbs sampling, and many asymmetric ones are computationally intensive. An effective alternative is Pitman Yor Processes (PYPs), also known as table indicator sampling, which exhibits faster convergence owing to dynamic memory requirements. Recent advancements have further improved PYP's computational speed, with only minimal space-time overhead compared to standard collapsed Gibbs Sampling. Integrated into the word-topic component, PYPs facilitate modeling of the topic model's word-generations and achieve quicker convergence through the use of collapsed Gibbs Sampling (Griffiths and Steyvers, 2004). Evaluation of topic models is another key component and we dive in more detail around it in the next Chapter.

Topic models have been employed in the social sciences, enabling the extraction of latent semantic structures from large text datasets and thereby providing insights into complex social phenomena across diverse disciplines such as sociology, public health, politics, media and the environment. In the field of sociology and anthropology, Mohr and Bogdanov (2013) applied LDA to unveil the underlying themes in sociological abstracts from the American Journal of Sociology, thus mapping out the intellectual landscape of the field over time. Paul and Dredze (2011) proposed employing topic models to capture public health trends based on Twitter data, whereas McCallum et al. (2007) used LDA to distil medical topics from patient

records, contributing towards targeted improvements in healthcare delivery. Monroe et al. (2008); Quinn et al. (2010) used topic models to draw insights from legislative speeches in US congress and distinguish language use among Democrats and Republicans, thus exposing profound disparities in political discourse. In a similar vein, Grimmer and Stewart (2013), used LDA and other adaptations to dissect the latent topics in U.S. Congressional speeches, aligning these topics to politicians' voting patterns and influence in policy making.

Looking at the domain of environment and climate change, Tvinnereim and Fløttum (2015) proposed the use of structured topic modelling (Roberts et al., 2014) to derive insights about public opinion from 2115 open-ended survey responses. Sleeman et al. (2017) presented dynamic topic modelling - a discrete topic model that maps topics over time (Blei and Lafferty, 2006c) to analyse the IPCC's reports and their extensive citations since 1990, accomplished by calculating cross-domain divergences between the citation domain and the report domain and clustering documents across domains, thereby uncovering the evolution of climate science and the influence of research trends on subsequent reports. Cheng et al. (2018) employed topic modelling on ecology, environment and poverty nexus from 4335 English language publication in the domain. Similarly, Rabitz et al. (2021) examined Lithuanian media's climate change discourse using the domestication framework and CTM for 583 news articles from 2017-2018. In Section 2.2, we discussed the language of Climate Change scepticism and briefly mentioned topic models. Boussalis and Coan (2016) used probabilistic topic modeling via LDA on 16,000 documents from 19 CCS organizations from 1998-2013. The generated topics covered sea and land impacts, scientific integrity, energy policies, and carbon trading politics, subsequently coded to separate science and policy frames.

2.5 Pre Trained Language models

In recent years, large transformer based Pretrained language models (PLMs) like BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019) have saturated natural language processing. PLMs are machine learning models that are trained on a large corpus and have learned a great deal about language patterns and real world knowledge, which can then be

fine tuned for specific downstream tasks like classification, summarization, generation, or question-answering. The use of pretraining in natural language processing (NLP) is a recent development, aided by the emergence of the powerful transformer architecture (Vaswani et al., 2017), however, the concept of pretraining has been applied in computer vision for at least a decade, with models such as ResNet (Huh et al., 2016; Targ et al., 2016; Yosinski et al., 2014).

2.5.1 Evolution of Pre Trained Language Models

Collobert and Weston (2008) were the first to propose pre training as a proxy to learn features rather than using rule-based approaches or feature engineering. The next line of work was learning word embeddings through word2vec (Mikolov et al., 2013a) and Global Vectors for Word Representation (GLoVE) (Pennington et al., 2014b) which are distributed representations of words in a lower dimension continuous vector space. Mikolov et al. (2013b) proposed 2 approaches to learn word2vec embeddings: cbow and skipg. cbow combines neighbouring words to predict a target word, while skip-gram uses the target word to predict neighbouring words. Building on it, Le and Mikolov (2014) introduced doc2vec embeddings to model whole documents or paragraphs through the means of dense vectors. GLoVE embeddings are created by leveraging global statistical information from a corpus to map words into a vector space and unlike word2vec which uses local context windows to learn vector representations, GLoVE constructs an explicit word-context or word co-occurrence matrix, allowing it to also capture global corpus-wide statistics. But the shortcoming of these models are; (1) not being able to tackle out of vocabulary words and; (2) the learned word embeddings being non-contextual meaning they are unable to capture context specific semantic differences. To tackle the issue of out of vocabulary words, Sennrich et al. (2016b) and Bojanowski et al. (2017) proposed byte pair encoding and fasttext to encode sub word units. We initialise downstream neural network models using them, so that most of the model parameters are not trained from scratch.

Moving on, the next innovation was the development of contextual representations, to address the shortcomings of static pre trained word embeddings which cannot capture

polysemy. The use of pretraining and fine-tuning in natural language processing (NLP) gained widespread popularity with the introduction of ELMo (Peters et al., 2018a) and ULMFiT (Howard and Ruder, 2018). Both models are based on the Long Short-Term Memory (LSTM) architecture (Hochreiter and Schmidhuber, 1997) and differ in their approach to pretraining. ULMFiT pretrains a three-layer LSTM on a language modeling task, while ELMo uses bidirectional LSTMs that perform language modeling in both the forward and backward directions. ULMFiT proposed fine-tuning the language model layer by layer for specific downstream tasks, and adding classifier layers on top of the language model which are also fine-tuned. In contrast after pretraining, ELMo simply extracts contextual word embeddings from it and adds these word embeddings to the downstream model, and the pretrained ELMO model is not updated when fine-tuning the downstream model for the task. The larger model size and larger pretraining dataset used by ELMo and ULMFiT allowed them to achieve competitive or improved performance on a range of tasks, demonstrating the effectiveness of pretraining language models at scale.

The introduction of the transformer architecture by Vaswani et al. (2017) paves the way for a new approach to building these pretrained models. The transformer model is based on a self-attention mechanism, which allows it to weigh the relevance of each neighbouring word in a sequence when encoding a particular word, thereby effectively capturing context regardless of the distance between words in the sequence. This is useful for a range of NLP problems and allows the model to learn more expressive representations, as multiple layers of self-attention can be used. Additionally, it also allows parallel computation, which significantly improves training efficiency compared to recurrent models like LSTMs, which have to process words sequentially, one word at a time. Furthermore, the architecture introduces positional encodings to account for the order of words in the sequence, which is not captured by self-attention mechanism.

Modern PLMs are built using the transformer architecture and can be broadly categorised into 3 classes; autogressive models also known as unidirectional models, Masked Language Models (MLMs) also known as bidirectional models, and encoder decoder based models (Min et al., 2021). An autoregressive language model is trained to predict the next word given

previous words, example models include GPT (Radford et al., 2018), GPT2(Radford et al., 2019) and GPT3 (Brown et al., 2020). The GPT family uses the transformer architecture as a decoder which consists of multiple transformer decoder layers with masked self-attention, that allows words to only attend to neighbouring words on the left (past), to accommodate the next-word prediction objective. The GPT family of models have been trained on increasingly larger amounts of text. GPT2 features 1.5 billion parameters compared to the original GPT with 117 million and thus has the capacity to model much more complex patterns in the data. One limitation of GPT2 is its inability to perform controlled generation for a specific domain.To address this issue, Keskar et al. (2019) introduced a conditional transformer language model capable of generating text based on control codes. These control codes guide the model's output, enabling the specification of various text aspects including style (e.g., review or tweet), sentiment (positive or negative), or subject matter (e.g., politics, sports, science). Building on this further, Dathathri et al. (2019) introduced a plug and play language model (PPLM), where the language model is a pretrained model similar to GPT2 but with smaller attribute models, which are trained to capture specific text characteristics, to control the generation of text, allowing for content with precise style, sentiment, or topic. Using an iterative refinement method, PPLMs can adapt to particular tasks or domains, leading to accurate and contextually relevant outputs. As a subsequent iterations of the GPT model, GPT3, introduced the remarkable capability of few-shot learning, reducing the reliance on fine-tuning. GPT3 leverages its generative design to address tasks either through prompt-based approaches or direct language generation, requiring minimal or no fine-tuning for specific tasks.

Masked language models (MLMs) are a type of model that predict a "masked" word in a sequence based on all of the other words in the sequence. Unlike autoregressive models, which process input in a single direction, MLMs consider context from both directions. During training, words are randomly masked or replaced with a special token [MASK] or a random token, which forces the model to consider context from both directions when making predictions. This allows a model to learn better contextual representations but loses the ability to perform language modelling in the standard left to right manner i.e. predicting the

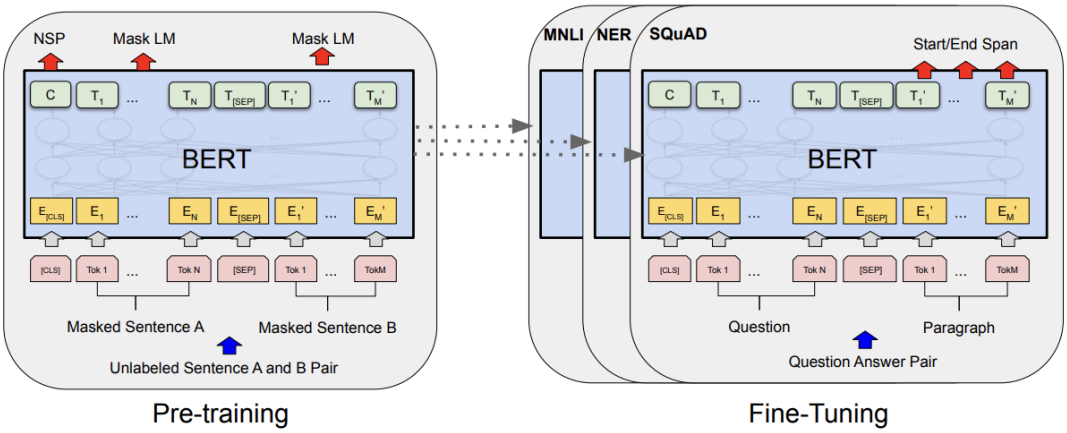


Figure 2.2: Architecture of BERT, reproduced from Devlin et al. (2019)

next word from previous *n* words. BERT (Devlin et al., 2019) is an example of an MLM that uses the transformer architecture and stacks multiple transformer layers to learn complex representations of words. One of the strengths of BERT is in transfer learning; where a model trained on one task is re-purposed on a second related task. In the context of BERT, once the model is trained on a large corpus of text, it can be fine-tuned with a smaller amount of task-specific data. This fine-tuning process involves additional training on a specific task, such as question answering or sentiment analysis, using a smaller labelled dataset. Both the pre training and finetuning architecure are shown in Figure 2.2. The pretraining paradigm is different to how static (word2vec) and contextual word embeddings (ELMo) are used in downstream tasks as it does not require a further downstream architecture or a separate neural network. For instance, to adapt BERT for a text classification task we only need to add an additional classification layer on top of BERT and fine-tune the model. There is a large family of models that were derived from BERT, including RoBERTa (Liu et al., 2019b), ALBERT (Lan et al., 2019), and DistilBERT (Sanh et al., 2019).

Liu et al. (2019b) introduced "Robustly Optimized BERT Pretraining Approach" (RoBERTa), with several modifications to the pretraining process of BERT. In contrast to the BERT static masking for its MLM training, RoBERTa uses dynamic masking, where the selection of masked tokens changes for each epoch of training, leading to more robust representations. Interestingly, RoBERTa removes the Next Sentence Prediction task from the pretraining

process, as the authors found that this task did not significantly contribute to the model's performance. Similarly, Lan et al. (2019) proposed ALBERT that implements cross-layer parameter sharing, which significantly reduces the model's size and increases training speed without a substantial loss in model performance. Instead of the Next Sentence Prediction (NSP) task used in BERT, ALBERT introduces a new pretraining task called Sentence Order Prediction (SOP), designed to better capture inter-sentence coherence.

To preserve the use of bidirectional context for masked word prediction and its ability to do left-to-right generation, Yang et al. (2019b) proposed XLNET. It leverages the permutation-based training of an autoregressive model, where all the possible factorization orders are considered (i.e. during training the model is trained to predict the next word using only left context words in some iterations and both left and right context words in other iterations), thereby maintaining the dependency among all input tokens. This methodology allows XLNet to learn bidirectional context like BERT while allowing the model to do generation. Sanh et al. (2019) proposed DistilBERT, a smaller, faster, and more efficient version of BERT, achieved through a process known as knowledge distillation which involves training a smaller model (student) to mimic the broad behaviour of a larger model (teacher). This idea of knowledge distillation has also been extended to other model types like GPT and RoBERTa. In the case of DistilBERT, the BERT model serves as the teacher. During training, the student model learns to approximate the teacher model's distribution of probabilities over the classes through the means of distillation loss computed as the cross-entropy between the softened output distributions of the teacher and student models. The softening is achieved by applying a temperature scaling factor to the output logits, which effectively smoothens the probability distribution over the classes. Both DistilBERT and ALBERT are lightweight variants of BERT, but how they achieve this is different.

The encoder-decoder model is based on the sequence-to-sequence framework, which uses a "text in, text out" process flow diverging from the uni-directional context understanding of models like GPT and the masked language modelling approach of BERT. Lewis et al. (2020a) introduced Bidirectional and Auto-Regressive Transformers (BART) that involves altering or corrupting input text with an arbitrary noising function, and then training the

model to reconstruct the original text. This process is bidirectional in that it has an encoder that considers word context from both preceding and following words, but also autoregressive in that it has a decoder that tries to regenerate the full output in a left-to-right manner, enhancing the model's ability and achieving superior performance over downstream tasks like generation, translation and summarization. In a smilar way, Raffel et al. (2019) devised Text-to-Text Transfer Transformer (T5), a model trained on denoising autoencoder objective, a methodology that draws inspiration from the aforementioned BART but with the distinction that every task in T5 is treated as a text generation problem. During pre-training, a portion of the input sequence is masked out, and the model is trained to predict the masked portion, thereby learning syntactic and semantic features from the text.

Gururangan et al. (2020) argued that while pretraining on large, diverse corpora can yield models with broad knowledge, these models may not perform optimally on specific tasks or domains due to the differences in data distribution. They proposed two methods of additional pretraining: Domain-Adaptive Pretraining (DAPT) and Task-Adaptive Pretraining (TAPT). DAPT involves further pretraining a language model on a corpus from the same domain as the target task whereas TAPT involved further pretraining over unlabeled task data. The rationale behind this approach is that by immersing the model in specific task data, it can more effectively learn the language patterns and nuances unique to that task, leading to enhanced performance. Taking this idea forward, multiple domain-specific model like FinBERT for financial data (Araci, 2019), LegalBERT for legal text (e.g., laws, court pleadings, contracts) (Chalkidis et al., 2020) and MedBERT for medical and electronic health records (Rasmy et al., 2021) have been developed. Webersinke et al. (2021) introduced one such model called ClimateBERT, a novel transformer-based language model, specifically tailored for climate-related texts. The model was pretrained on a corpus of over 2 million paragraphs derived from diverse sources such as news, corporate disclosures, and scientific articles. The training approach employed is a domain-adaptive pretraining strategy, as proposed by Gururangan et al. (2020). This three-stage process begins with pretraining on a general domain using the distilled version of RoBERTa (Liu et al., 2019b). The second stage involves further pretraining on the target domain, i.e., climate-related texts, with the application of

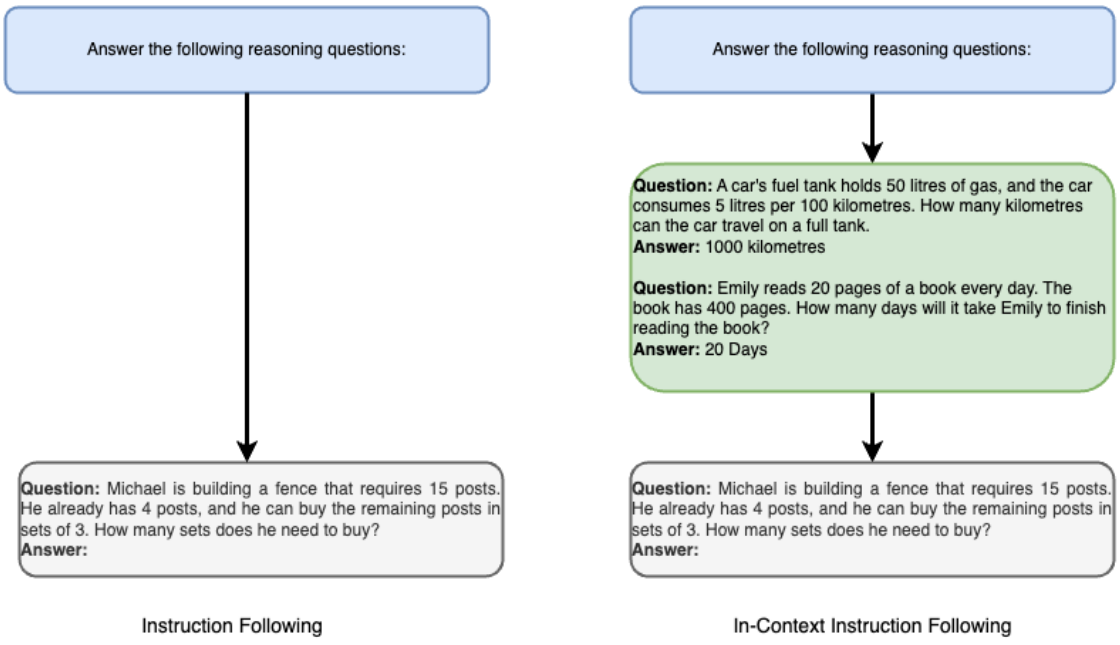


Figure 2.3: An example of Instruction Following

two sample strategies: SIM-SELECT and DIV-SELECT. These strategies involve the use of a subset of the corpus that is either similar to or diverse from the samples in the downstream task, respectively. The final stage involves training the model on specific downstream NLP tasks such as fact-checking, sentiment analysis, and classification.

2.5.2 Large Language Models

The evolution of PLMs led to the development of Large Language Models (LLMs), which are essentially scaled-up versions of PLMs, distinguished by their substantial size and comprehensive capabilities. LLMs are adept at understanding and generating text that closely mirrors human language, excelling in a variety of tasks such as translation, question-answering, summarization, and text generation. A notable emergent property of these scaled-up LLMs is in-context learning, as seen in GPT-3 (Brown et al., 2020). In this paradigm, input-output pairs (few shot) are provided as demonstrations, serving as explicit examples to guide the model in learning the underlying function or pattern of the desired task.

Another significant emergent property arising from the scaling of LLMs is their ability to understand and follow natural language instructions, a capability referred to as "instruction following". This can be achieved through 2 notable methods: (1) direct supervised instruction tuning, as seen in models like FLAN T5 (Chung et al., 2022), and; (2) reinforcement learning from human feedback (RLHF), as employed in models like ChatGPT.[12] Instruction following was initially conceptualized as a zero-shot approach, where the model was expected to understand and correctly respond to natural language instructions without prior examples. Subsequently, there has been an effort to integrate in-context learning (few-shot) with instruction following, further enhancing the model's adaptability and effectiveness, as illustrated in Figure 2.3.

Instruction following involves fine-tuning a pre-trained model on a dataset specifically curated to contain a variety of tasks formatted as instructions. Thus the model learns to follow instructions directly from these examples and becomes adept at interpreting and executing a wide array of tasks based on direct instructions. An instruction-formatted instance broadly consists of a 2 components: (1) task description (called an instruction) like "Translate the following sentence into French" or "Summarize the below paragraph" and; (2) an input with its corresponding output. Numerous models, employing instruction tuning in varying configurations, have demonstrated promising results. For instance, Flan T5 (Chung et al., 2022), a derivative of T5, focuses on multi-lingual tuning for translation and Question Answering tasks. BLOOM (Scao et al., 2022) excels in tasks like reading comprehension, numerical reasoning, and common sense logic, while GPT-3 (Brown et al., 2020) and OPT (Iyer et al., 2022) demonstrates proficiency in summarization, question answering, and translation. Models like Codex (Chen et al., 2021) and Palm (Chowdhery et al., 2022) have showcased their strength in translating natural language to code and handling multi-modal instructions comprising text, images, and code, respectively. A critical facet of instruction tuning datasets is the diversity of instructions, generally missing in large public datasets. To address this Ouyang et al. (2022), proposed InstructGPT which involves leveraging queries

[12]https://openai.com/blog/chatgpt

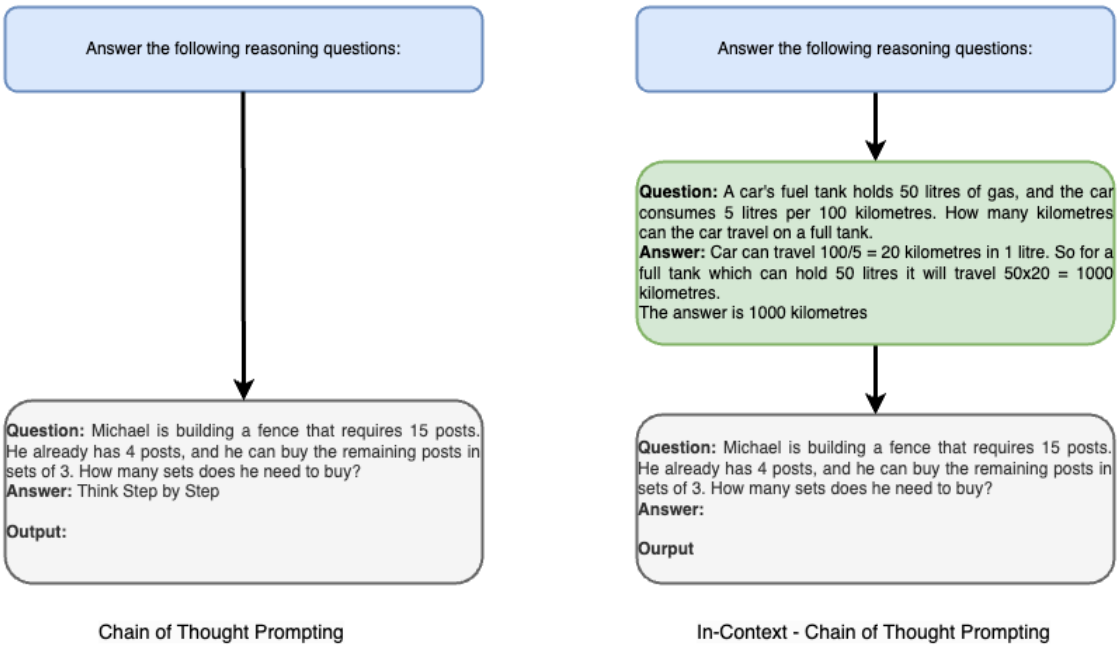


Figure 2.4: An example of Chain Of Thought Prompting

submitted by real users to the OpenAI API which are generally phrased in natural language, serve as “task descriptors”, and human labelers are used to generate corresponding answers.

On the other hand, RLHF involves an initial phase of supervised fine-tuning, followed by iterations where the model’s responses are evaluated by human trainers (Bai et al., 2022; Christiano et al., 2017). Feedback from these evaluations is used to further train the reward model, which serves to help align — through reinforcement learning the LLM’s responses more closely with human preferences and instructions, thereby making it to more effective at following instructions that is satisfactory to human users, improving in areas like conversation quality, relevance, and safety.

While instruction following has proven effective, recent research has pivoted towards enhancing prompting techniques for more complex tasks. Recognizing the limitations of simple reasoning research has been focussed on decomposing tasks into smaller, more manageable sub-problems by guiding the model through a series of steps to solve the overall, more complex task. Kojima et al. (2022) and Wei et al. (2022) introduced a strategy known as the “Chain of Thought" (CoT) which encompasses a set of intermediate instructions

designed to depict reasoning or rationales, culminating in the final answer. Kojima et al. (2022) demonstrated the effectiveness of CoT in a zero shot setting by simply adding “Let’s think step by step” to the original prompt whereas Wei et al. (2022) applied CoT with a few-shot examples, effectively bringing in in-context learning, as depicted in Figure 2.4. Yao et al. (2023) extended CoT by proposing “tree of thought”(ToT) which explores multiple reasoning pathways at each step of the problem-solving process. It begins by breaking down the problem into various thought steps and then generates multiple lines of reasoning for each, effectively creating a branching tree structure.

2.6 Misinformation Detection

In this section we will first review the literature surrounding general misinformation, fact checking and different terminologies connected to it. We discuss the development of datasets in NLP and simultaneously delve into the stylistic components that characterize misinformation, and the diverse methodologies that have been used for its detection. As we progress, our focus shifts towards the domain of climate change, casting a spotlight on fact checking in this context.

The umbrella term “misinformation" refers to any piece of information that is false, inaccurate, or misleading, regardless of the intention behind its creation or dissemination (Kuklinski et al., 2000; Vraga and Bode, 2020). Misinformation though a general phrase can materialize in myriad of forms, often overlapping in meaning. For instance *disinformation*, a subtype of misinformation, specifically signifies intentionally fabricated and circulated false information with the explicit objective to deceive or misguide audiences (Wu et al., 2019). *Fake news* is another variant of misinformation, referring to entirely falsified information, often sensationalized, that masquerades as legitimate news reporting with motivation ranging from the generation of advertising revenue to the manipulation of public sentiment (Allcott and Gentzkow, 2017). In the sphere of public discourse, *false claims* are characterized by inaccurate or erroneous assertions or statements made by individuals or entities which can be intentional or unintentional. *Propaganda* is the deliberate propagation of information, often

biased or misleading, aimed at influencing public opinion or obscuring the truth with tactics employed often involve sophisticated manipulation and persuasion techniques, ranging from emotional appeals, sensationalism and false dilemmas to demonizing perceived adversaries (Jowett and O'donnell, 2018). Similarly, *rumours* are unverified pieces of information that circulate widely without substantiated evidence, and are particularly widespread on social media platforms (Qazvinian et al., 2011). An extension to this is *troll content*, framed as disruptive or deceptive online behaviour that ranges from harmless pranks to more serious activities, including the dissemination of hate speech and disinformation, with the intent of inciting discord or confusion within online communities (Buckels et al., 2014).

Broadly in the realm of this thesis, we discuss combating misinformation through the tasks of fake news detection and claim verification. Fake news detection, involving classification of news articles as "fake" or "real", utilises inputs like article text and associated metadata, with outputs ranging from binary labels to probabilistic scores of reflecting the likelihood of the news being fabricated. In contrast, claim verification assesses the veracity of specific claims using inputs like the claim and contextual information, with more granular output categories reflecting the degree of truthfulness (Hassan et al., 2015; Thorne et al., 2018).

Early research on misinformation focused on fake news detection (Vlachos and Riedel, 2014), claim/stance verification (Ferreira and Vlachos, 2016), and propaganda detection (Barrón-Cedeno et al., 2019; Da San Martino et al., 2019). The lack of annotated fake news data spurred the creation of misinformation datasets. The first public dataset for fake news detection (Vlachos and Riedel, 2014) and claim/stance verification (Ferreira and Vlachos, 2016) were relatively small with 221 and 300 instances, respectively. More recently, larger datasets have been developed, such as LIAR (Wang, 2017) and FEVER (Thorne et al., 2018). LIAR comprises of 12.8K manually annotated short statements collected from PolitiFact and labelled with 6 levels of veracity which range from "true", "mostly true", "half true", "mostly false", "false" to "pants on fire", providing a nuanced gradation of truthfulness and falsity. Additionally, it also includes metadata related to the speaker, the speaker's title, the state, the party affiliation, where the statement was made, and the subject of the statement. The Fact Extraction and VERification (FEVER) dataset (Thorne et al., 2018), composed of

roughly 185,000 claims, is a dataset generated from Wikipedia with "supported", "refuted" and "not enough info" labels. In contrast to LIAR, FEVER directly links claim truthfulness to extractable evidence from Wikipedia, offering a valuable resource for claim verification and evidence extraction. Extending the focus to full articles rather than just individual statements, Shu et al. (2018, 2017) constructed FakeNewsNet which has news content mixed with social context and additional spatio temporal information. The dataset includes two subsets: one from PolitiFact and one from GossipCop, both labeled as either `Real` or `Fake`.

Style broadly refers to the properties of a sentence beyond its content or meaning (Pennebaker and King, 1999), and stylistic variation plays an important role in the identification of misinformation. Biyani et al. (2016) studied stylistic aspects of clickbait and formalised it into 8 different categories ranging from exaggeration to teasing, and proposed a clickbait classifier based on novel informality features. Similarly, Kumar et al. (2016), examined the unique linguistic characteristics of hoax documents in Wikipedia and built a classifier using a range of hand-engineered features. Rashkin et al. (2017) proposed using stylistic lexicons (e.g. Linguistic Inquiry and Word Count (LIWC)), subjective words, and intensifying lexicons for fact checking, and demonstrated that words used to exaggerate such as superlatives, subjectives, and modal adverbs are prominent in fake news, whereas trusted sources are dominated by assertive words. Horne and Adali (2017) demonstrated that the style of fake news more closely resembles satire than real news, with title structure and the use of proper nouns being significant differentiators. Additionally, stylometry and psychological features also play a crucial role in distinguishing between fake and real news. On similar lines, Przybyla (2020) explored the task of low credibility online documents based on their writing style. They collected a corpus of over 103,000 documents from 223 online sources, and created two new classifier, a neural network and a stylometric model which focussed on sensational and affective vocabulary present in fake news. Wang (2017) experimented with detecting fake news using metadata features with convolutional neural networks adapted for text (Kim, 2014).

With the emergence of transformer-based models (Vaswani et al., 2017), a new line of research developed, focusing on their use on using them over handcrafted features and

shallow learning models for fact checking. Jwa et al. (2019) proposed an automatic fake news detection model which employs BERT but with weighted cross entropy to account for class imbalance, with additional pretraining using the CNN and Daily News Mail dataset (Nallapati et al., 2016). Liu et al. (2019a) developed a two-stage model and treated the task as a fine-grained multi-classification task. In the first stage, BERT is used to extract features, which are then enhanced with additional metadata and in the second stage, two similar sub-models are employed to separately identify labels of different granularities. Vijjali et al. (2020) experimented with fact checking for COVID-19 claims using a 2 layer process where first stage acts as a retriever to fetch explanations for candidate claims, and the second stage is modelled as text entailment where the claim and explanations are assessed for veracity. Kaliyar et al. (2021) suggested a method that integrates various parallel blocks of a single-layer deep convolutional neural network, each with distinct kernel sizes and filters, in conjunction with BERT. Moving on to different transformer architectures, Raza and Ding (2022) introduced FND-NS (Fake News Detection through News content and Social context) which adapted BART for the task of detecting fake news, taking into account both the content of the news and its associated social contexts.

Research on fake news and propaganda has primarily operated at the article level, and focused on binary detection (presence vs. absence) (Barrón-Cedeno et al., 2019; Rashkin et al., 2017). Da San Martino et al. (2019) argued for the need for granularity in propaganda detection, both in terms of propaganda sub-types and fragment-level detection. To facilitate this, the authors developed a corpus of news articles that are manually annotated at the fragment level with eighteen different propaganda techniques, and designed a novel multi-granularity neural network. In a similar vein, Nakamura et al. (2020) proposed Fakeddit, a multimodal dataset designed specifically for fine-grained fake news detection. The dataset includes over 1 million samples from various categories of fake news, encompassing text and image data, metadata, and comment data and developed hybrid text+image models for multiple variations of classification aiming to distinguish between misleading, manipulated, and completely false content.

Although articles with misinformation are predominantly human-written, the recent emergence of large pre-trained language models means they can now be automatically generated. To protect against the threat of fake news generation by such models at scale, Zellers et al. (2019) proposed GROVER, a conditional transformer based language model with a semi-supervised discriminator to defend against neural fake news.

We now review misinformation literature related to climate change. Several studies have analysed the discourse around climate change. Diggelmann et al. (2020) formalised the task by introducing CLIMATE-FEVER as a veracity prediction task, as an extension of FEVER for fact checking climate change fact checking claims. In Chapter 5, we use CLIMATE-FEVER, and describe it in more detail here. The pipeline to construct CLIMATE-FEVER was similar to FEVER and consists of (1) collecting climate change related claims scraped from the web using seed keywords related to climate change, which were later verified with the help of climate scientist and (2) constructing corresponding evidence retrieval sentences. The evidence retrieval pipeline consists of 2 parts: (1) evidence candidate retrieval system (ECRS) and; (2) evidence candidate labelling (ECL). In the case of ECRS, the process begins by retrieving relevant documents (related to a claim) from Wikipedia using an information retrieval method as BM25 (Robertson et al., 1995). From these documents, the top 100 sentences are selected using task-specific sentence embeddings in an average-pooled Siamese setting (Reimers and Gurevych, 2019a) and finally, these sentences are re-ranked using a pre-trained Albert model (Lan et al., 2019b) applied to the FEVER dataset, providing a relevance score for each claim-evidence pair.[13] The top 5 ranked sentences comprise the final evidence set. In the ECL step, each claim with its evidence is given to 5 annotators for labelling with the classes 'supported', 'refuted' and 'not enough info'. For each claim, a micro verdict is determined based on a majority vote (out of 5 annotators) for each claim evidence pair and a macro label is computed as an aggregation over all 5 micro verdicts (each claim has 5 evidence items, hence 5 micro verdicts).

[13]This is done over 2 classes to predict a relevance score i.e. evidence and non-evidence, with evidence being the corresponding evidence sentences from FEVER and non-evidence being random sentences from Wikipedia dump.

In terms of explainable fact checking in climate change, Atanasova et al. (2020) used DistilBERT in a multitask setting and performed the joint task of summarisation and classification of the veracity of the claim. Stammbach and Ash (2020) experimented with GPT3's (Brown et al., 2020) few shot learning capabilities to generate fact-checking explanations. As mentioned in Section 2.5, Webersinke et al. (2021) introduced ClimateBERT, which is a transformer based model based on BERT which was pretrained on news, scientific articles, and corporate disclosures related to the domain of climate change. They employed this system to experiment with the downstream tasks of classification (e,g to check if a paragraph is climate related or not), sentiment analysis (e.g. to assess a report being climate negative risk vs positive opportunity), and fact checking (e.g. CLIMATE-FEVER). Chillrud and McKeown (2021) worked on the task of climate change fact checking by adding unsupervised data augmentation (Xie et al., 2020b) to the usual fact checking pipeline and applied it to CLIMATE-FEVER (Diggelmann et al., 2020). In this setting, data augmentation is performed though back translation of a data point i.e. a claim is translated from English to German and then back from German to English thereby adding noise through paraphrase. The loss function comprises 2 parts: (1) standard cross entropy loss on labelled data between predicted output and the gold label; and (2) consistency loss between the model prediction (over unlabelled data) of original example and its back translated counterpart.

2.7 Summary

In this chapter, we presented a review of the literature through two lenses: social sciences and NLP. From the social science perspective, we provided an overview of the history of the climate change debate, the climate change skepticism movement, and the narrative and stylistic elements utilized. We also presented literature on the concepts of framing and neutralization and their role in shaping narratives. In the second aspect, we detailed NLP approaches, including topic modelling, pre-trained language models and its various architectures, large language models and their emergent properties like instruction following. Finally, we discussed misinformation and claim verification. In the next four chapters, we

will present our research on optimizing topic models for better representation at the document level (Chapter 3), detecting climate change scepticism articles (Chapter 4), understanding and detecting the framing and neutralization tactics used in constructing climate change scepticism narratives (Chapter 5), and fact-checking the veracity of claims, elucidating reasons for potential inaccuracies (Chapter 6).

Chapter 3

Topic Intrusion for Automatic Topic Model Evaluation

This chapter builds on the paper:

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "Topic intrusion for automatic topic model evaluation." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 844-849. 2018.

3.1 Introduction

Topic models such as Latent Dirichlet allocation (Blei et al., 2003) are capable of simultaneously learning latent topics in the form of multinomial distributions over words and the allocation of topics to individual documents in the form of multinomial distributions over topics. They are highly effective in extracting themes and concepts and helping users visualize document collections. As previously discussed in Chapter 2, topic models have been applied in various social science domains like sociology (Mohr and Bogdanov, 2013), health (Paul and Dredze, 2011), politics (Monroe et al., 2008; Quinn et al., 2010), and environmental studies (Cheng et al., 2018; Tvinnereim and Fløttum, 2015) to analyze large datasets but it isn't always clear how to build an optimal models. Traditionally, due to the unsupervised

nature of topic models, they have been optimised using intrinsic evaluation metrics like perplexity but it has been shown to correlate poorly with direct human assessment of topic model quality (Chang et al., 2009), motivating the need for automatic topic model evaluation methods which emulate human assessment.

Increasingly, topic coherence is being used as a task-independent evaluation method (Aletras and Stevenson, 2013; Lau et al., 2014; Mimno et al., 2011; Newman et al., 2010; Röder et al., 2015) but this approach often ignores the topic allocations to individual documents. Topic models should not only produce coherent topics but should also capture general concepts from our document collection. Bhatia et al. (2017) showed that topic coherence as a standalone evaluation can be misleading. They demonstrated this with an adversarial topic model that generates highly coherent topics, which, while individually meaningful, collectively provide limited insights into the overall content of the document collection but provided no evaluation at the document level. Our work aims to address this gap.

In the previous chapter we presented literature about topic model, their formulation, types of topic models and a few applications but largely ignored a discussion around evaluation. We commence this chapter by giving an overview of evaluation of topic models both in terms of intrinsic evaluation through the means of perplexity and extrinsic evaluation through the lens of topic coherence. To be able to compare and distinguish between good and bad topic models and have a document level evaluation we will need different collections and topic models. To this end, next we introduce the 2 collections namely (APNEWS and BNC) and 5 different topic models.

Following this, we detail the task of topic intrusion, which forms the core of our evaluation. We base it on the topic intrusion setup first introduced by Chang et al. (2009) where users are presented with a document, a set of associated topics (from a topic model) and an intruder topic, and are tasked to find the intruder. Then, employing this setup we detail the process of collecting human judgements which is helpful for manual evaluation of our methodology. Collecting annotations can be both time consuming and expensive, thus the need for automating this task. Thus, we follow it up by introducing a convolutional neural network approach to automate the this task and provide additional analysis of document-

level evaluation via mean-absolute-error with the help of collected annotations. Finally, we propose a new metric to measure the performance of the system, use it to optimise topic models and rank topics produced by topic models to get an overview of topics generated.

In Chapter 1, we mentioned that the research was conducted over a span of 5.5 years, with the timeline of the work in this chapter corresponding to the first half of 2018. Consequently, within the scope of this chapter, we utilize an architecture based on Convolutional Neural Networks integrated with text embeddings. This choice was made because, at the time, a combination of these models along with Recurrent Neural Networks was widely employed in the field.

3.2 Background

Chang et al. (2009) introduced two methods of evaluating the quality of topic models using human judgements, one at a topic level and the other at the document level through the means of word and topic intrusion tasks, respectively, to assess topic models. The authors showed a low correlation between the measure of perplexity commonly used in topic modelling and direct human evaluations of topic model quality. These two methods take the form of “intruder” tasks, where participants are asked to identify an intruder topic word or an intruder topic for a given topic or document. Specifically, in the word intrusion task, an intruder word is added to a list of the top 5 topic words and participants are asked to identify which word does not belong. Similarly, in the topic intrusion task, a document and four topics are presented — three corresponding to the document and one being an intruder topic - and participants are asked to identify the intruder topic. The idea is that for high quality topics should it be easier to spot the "intruder" in these tasks. Since then, various automatic measures to assess topics have been proposed (Aletras and Stevenson, 2013; Mimno et al., 2011; Newman et al., 2010).

Newman et al. (2010) suggested a direct way to measure topic coherence through manual annotation, where the annotators rate topics on a 3-point ordinal scale. To automate the estimation of topic coherence, they compute a score which is an average of a similarity

measure for all pairs of topic words. They tested with different similarity measures such as association measures, Wikipedia-based measures, WordNet-based measures and a search engine based measure and found that the best results were achieved by calculating pointwise mutual information (PMI) scores over English Wikipedia. In a similar vein, Mimno et al. (2011) also proposed coherence as an evaluation metric but with two differences: (1) they employed conditional probability instead of PMI; and (2) they used the training set of the topic model to compute co-occurrence counts instead of Wikipedia.

Following this, Lau et al. (2014) conducted a comprehensive comparison and analysis of various approaches to topic coherence and proposed a method grounded in Normalised Pointwise Mutual Information (NPMI). Additionally, they introduced an automated procedure for assessing the human-interpretability of topics, leveraging the concept of word intrusion, a task that had traditionally been dependent on manual annotation. Their methodology incorporated word association features such as PMI, NPMI, CP1, CP2 which were trained in a learning-to-rank Support Vector Regression model (Joachims, 2006). Lau et al. (2014) investigated the relation between coherence scored using word intrusion vs. direct estimation (i.e. computing similarity between word pairs in a topic) and found high correlation. Röder et al. (2015) proposed a framework that allows the construction of existing word-based coherence measures, as well as new ones, by combining elementary components. They conducted a systematic search of the space of coherence measures using all publicly available topic relevance data for evaluation and showed that new combinations of components outperform existing measures with respect to correlation to human ratings.

Bhatia et al. (2017) revisited the topic intrusion task of Chang et al. (2009), and explored its viability as an alternative task-independent approach for topic model evaluation. They tested a number of topic models and found that there can be large discrepancies between conventional topic coherence measures and topic intrusion results, suggesting that topics can be individually coherent but poor descriptors of the documents. In addition, they proposed a method to automate the topic intrusion task by training a support vector regression model based on information retrieval (IR) and word co-occurrence features to predict the intruder topic and reported encouraging correlation levels with human judgements for model-level

evaluation thereby demonstrating that topics learnt by the topic model are relevant to the document.

Although Bhatia et al. (2017) is able to distinguish between good and bad topic models (at the model-level), their work provided no evaluation at the document level other than the observation that 'there are still slight disparities between human annotators and the automated method in intruder topic selection". Our work extends Bhatia et al. (2017) as follows: (1) we improve the results based on a novel neural model and provide additional analysis of document level evaluation via mean-absolute-error; and (2) we propose a new metric to measure the performance of the system thereby optimizing the topic model on topic-document allocations. In this work we will using Bhatia et al. (2017) as a comparison for our experiments.

3.3 Datasets and Topic Models

In this section we will describe the datasets and topic models used for our experiments. In terms of dataset we conduct our experiments using the datasets employed by Bhatia et al. (2017):

1. APNEWS which is a collection of Associated Press news articles from 2009 to 2016. We randomly sampled 50K documents.

2. British National Corpus ("BNC": Burnard (1995)) which is a collection made up of excerpts from diverse sources such as journals, books, letters, and articles. We randomly sampled 15K documents.

Similarly to Bhatia et al. (2017); Chang et al. (2009), we base our analysis on a representative selection of topic models, each of which we train over APNEWS and BNC to generate 100 topics: For the topic models we experiment with the following: standard Latent Dirichlet Allocation (`lda`: Blei et al. (2003)), correlated topic model (`ctm`: Blei and Lafferty (2006b)), non-parametric topic model (`hca`: Buntine and Mishra (2014)), neural topic model (`ntm`:

Cao et al. (2015)), and an adversarial topic model (`cluster`: Bhatia et al. (2017)). We reviewed these topic models in Section 2.4 and will briefly describe them here as well.

- LDA is one of the most established topic models and uses a symmetric Dirichlet prior to model both the mix of topics within a document and the mix of words within a topic. We use Mallet's implementation of LDA, which includes some enhancements to the basic LDA model such as the use of an asymmetric-symmetric prior.[1]

- CTM is an extension of LDA which replaces the dirichlet prior with a logistic normal prior over topic proportions. It is designed to model the correlations between different topics and minimize overlap in topic content.

- HCA is an extension of LDA that has the capability to capture word burstiness (Doyle and Elkan, 2009), which refers to the tendency for a word to be more likely to be generated again after it has been seen recently, and is modelled through the means of a Pitman-Yor Process (Chen et al., 2011).

- NTM uses a neural network architecture where the topic-word multinomials are modelled as a look-up layer of words and the topic-document multinomials are modelled as a look-up layer of documents. The output layer of the network is then calculated as the dot product of these two vectors.

- Cluster is an adversarial topic model in the sense that it is designed to produce topics that are coherent but poor descriptors of documents. It is a topic model that generates highly coherent topics but simple topic allocations by using pre-trained word2vec vectors, with k-means clustering to create word clusters. The multinomial topic distribution is generated by taking the cosine distance to the cluster centroid and linearly normalizing it across the words. To generate a topic allocation for a document, it calculates the document's representation by computing the mean of the word2vec vectors of its content words and then calculates the similarity of the document to each cluster by cosine similarity and normalizes it linearly.

[1] https://mimno.github.io/Mallet/

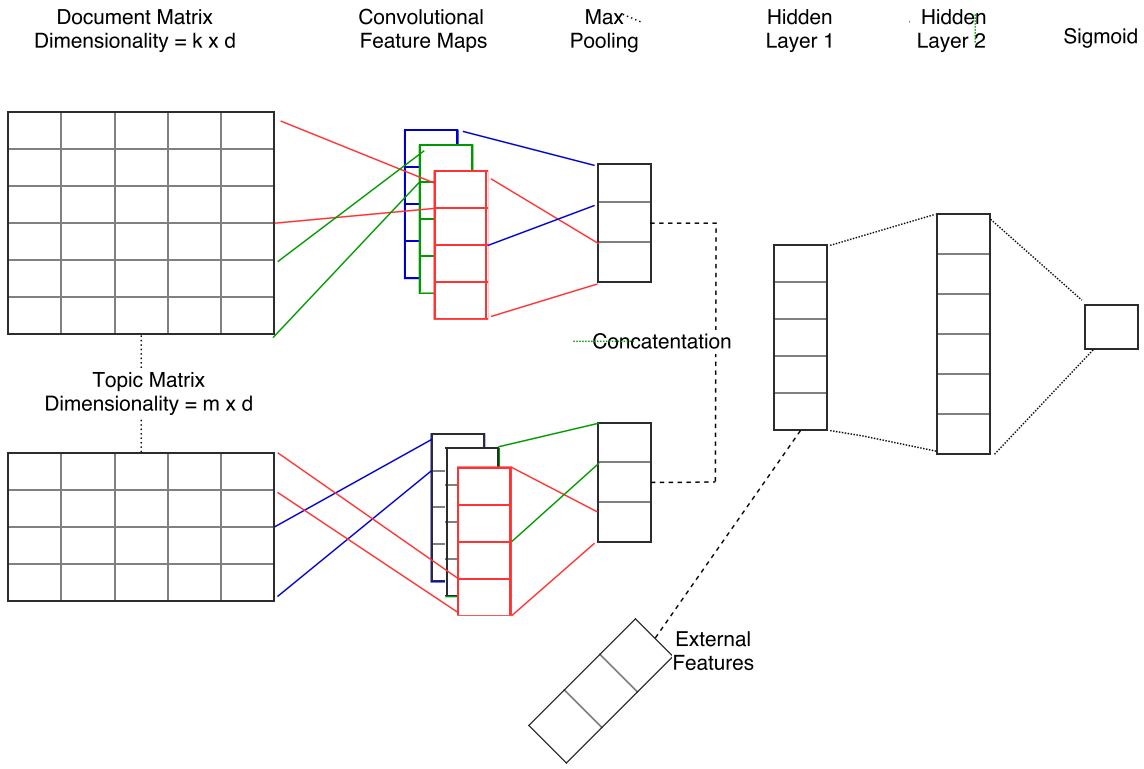


Figure 3.1: Architecture diagram of our method

3.4 Methodology

In this section, we first briefly describe the topic intrusion task, followed by the the human annotation collection process and finally propose an improved methodology to automate it.

3.4.1 Task

Chang et al. (2009) first proposed the topic intrusion task with the aim of assessing whether topics associated with a document capture its content. In this task, an annotator is presented with a document along with its top-3 highest probability topics and a low probability unrelated intruder topic, and are asked to identify the outlier intruder topic with the expectation being that a better topic model will make it easier to identify the intruder topic. Bhatia et al. (2017) incorporated additional constraints: (1) it must be a low probability topic for the document in question; and (2) the intruder topic has to have high probability for at least one other

document. Their rationale behind the first constraint is that it ensures that the intruder topic is unrelated to the target document while the second ensures that the intruder topic is coherent and interpretable. Each topic is represented by its top 10 most probable words, and the first three sentences of the target document are provided, with the option to view more of the document if needed.

3.4.2 Human Judgements

To assess our methodology, we need human annotations for the topic intrusion task. We collect human judgements using Amazon Mechanical Turk. Each HIT is comprised of 5 documents, and each document is paired with 4 topics (3 real and 1 intruder). To control for annotation quality, an additional document–topics pair is inserted as part of the HIT. The control item's intruder topic is generated by randomly sampling words from the corpus vocabulary. To pass the quality control, an annotator has to select the correct intruder topic; they are filtered out if their pass rate over all controls is < 0.6. We fixed the threshold to 0.6 based on preliminary experiments. We found that it was a challenging task, and this value provides quality without filtering out most of the workers. To better understand the task we present screenshot of the task for a single HIT in Figure 3.2, which shows the instructions for the task with an example for the annotators followed by the the task to be annotated. Looking at the task, we see for the second document that the intruder topic is the first topic as it is related to alcohol and not remotely related to the document (which is about traffic accident). For the first document it is a bit more difficult. Note, we only see the first 2 document-topic pair for this HIT but as mentioned above, 5 such document-topic pairs comprise a HIT.

Each HIT is judged by 10 workers. We collect additional annotations by releasing the task internally to a small number of local workers. We needed to carry out some annotations internally to make sure that each HIT had at least 4 annotations. The average number of internal annotations per HIT was approximately 1.6. For each topic model, we collected annotations for 100 documents on 2 corpora and 5 topic models. Hence in total we annotated 1000 document-topic combinations (5 topic models $\times$ 100 documents $\times$ 2 collections $=$

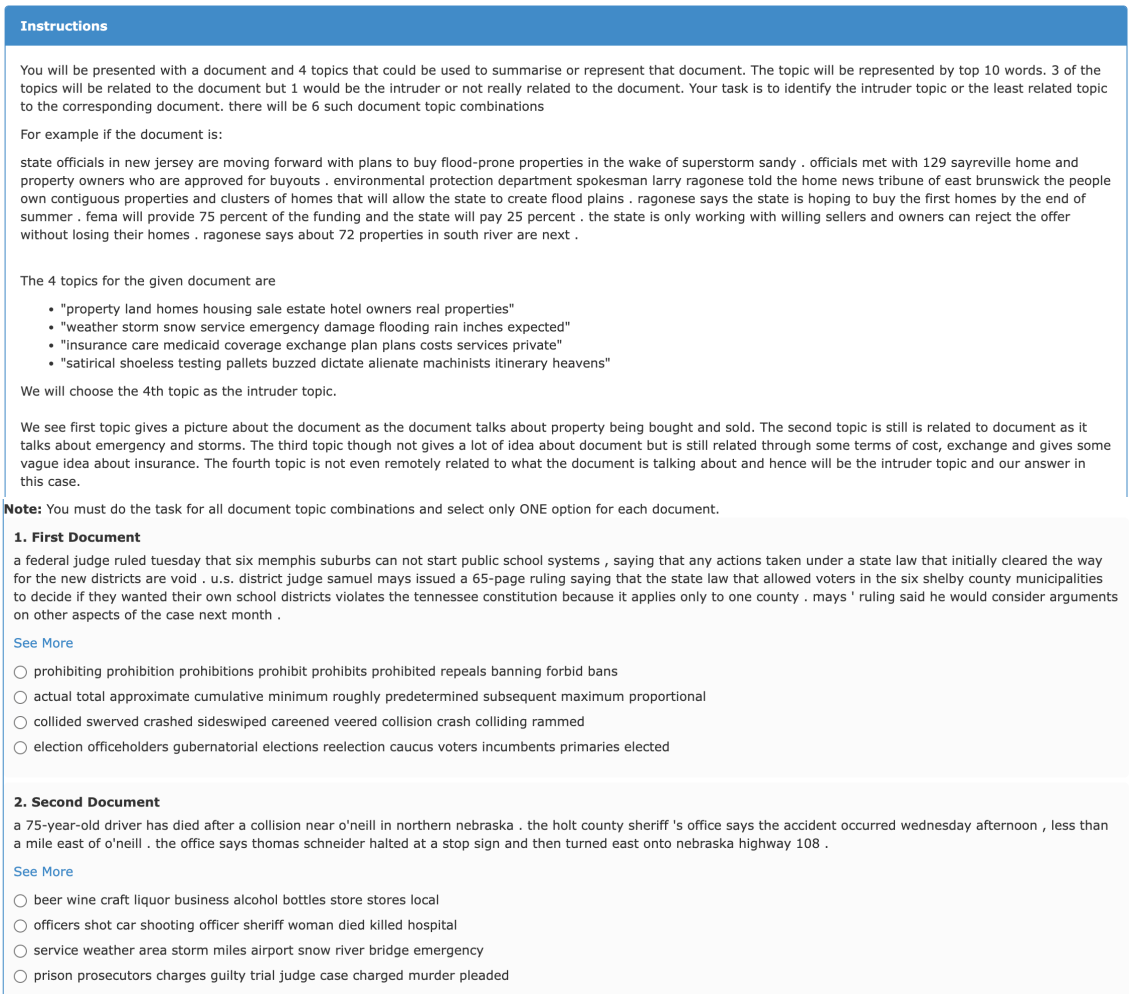


Figure 3.2: Screenshot of a HIT

1000). After filtering and including internal judgements, we have an average of 6.7 and 6.9 annotations for APNEWS and BNC, respectively.

3.4.3 Intruder Topic Detection

We propose a neural network model to automatically predict intruder topics. Our model is inspired by Severyn and Moschitti (2015), where they combine a learn-to-rank deep learning architecture in an IR setting to rank the documents for a given query. We adapt it to our topic intrusion task by ranking topics for a given document. Our task takes the form of a document d_i with corresponding topics $T_i = \{t_i^1, t_i^2, t_i^3, t_i^4\}$, where 3 topics are real and 1 is

the intruder. The topic set T_i has labels $Y_i = \{y_i^1, y_i^2, y_i^3, y_i^4\}$ ("1" denotes the intruder topic, or "0" otherwise). We train using a pointwise ranking approach, where training examples are triples of (d_i, t_i^j, y_i^j) — essentially the task is formulated as a binary classification problem.

The architecture of our network is given in Figure 3.1 and is based on the convolutional neural network architecture proposed by Kim (2014) and Severyn and Moschitti (2015). The input to our model is a document–topic pair, represented as a sequence of words. Let $x_{d_i} \in \mathbb{R}^u$ and $x_{t_1}\mathbb{R}^u$ be u dimension word vector for the i-th word in the document and topic respectively. A document of length k ($k =$ document length) and topic of length m ($m =$ number of topic words) can be given by Equation 3.1 where $\oplus$ is a concatenation operation.

$$\begin{aligned} x_{d_{1:k}} &= x_{d_1} \oplus x_{d_2} \oplus \ldots \oplus x_{d_k} \\ x_{t_{1:m}} &= x_{t_1} \oplus x_{t_2} \oplus \ldots \oplus x_{t_m} \end{aligned} \tag{3.1}$$

This can also be expressed in terms of embeddings, via embedding matrix $W \in \mathbb{R}^{|V| \times u}$, where V is the vocabulary and u the dimensionality of the embeddings to give document embeddings $E_d \in \mathbb{R}^{k \times u}$ ($k =$ document length) and topic embeddings $E_t \in \mathbb{R}^{m \times u}$ ($m =$ number of topic words). A convolution operation can be applied on a window of h words through a kernel $w \in \mathbb{R}^{hu}$ to produce feature maps for both the document and topic (Kim, 2014; Severyn and Moschitti, 2015). For example a feature c_{d_i} and c_{t_i} is generated for document and topic respectively as shown in Equation 3.2.

$$\begin{aligned} c_{d_i} &= f(w \cdot E_{d_{i:i+h-1}} + b), \quad i = 1, \ldots, k-h+1 \\ c_{t_i} &= f(w \cdot E_{t_{i:i+h-1}} + b), \quad i = 1, \ldots, m-h+1 \end{aligned} \tag{3.2}$$

where f is an activation function like RELU and $b \in \mathbb{R}$ is a bias term. The kernel can be applied to every possible windows $\{E_{d_{1:h}}, E_{d_{2:h+1}} \ldots E_{d_{k-h+1:k}}\}$ to generate feature maps as

given in Equation 3.3 where $c_d \in \mathbb{R}^{k-h+1}$ and $C_t \in \mathbb{R}^{m-h+1}$.

$$\begin{aligned} C_d &= [c_{d_1}, c_{d_2}, ..., c_{d_{k-h+1}}] \\ C_t &= [c_{t_1}, c_{t_2}, ..., c_{t_{m-h+1}}] \end{aligned} \tag{3.3}$$

It is followed by a max pooling operation.

$$\begin{aligned} \hat{C}_d &= \max C_d, \quad (7) \\ \hat{C}_t &= \max C_t, \quad (8) \end{aligned} \tag{3.4}$$

The convolution and max pooling operation is performed using feature maps of varying size to produce a constant-length vector P in Equation 3.5

$$P = [\hat{C}d_1; \hat{C}d_2; \ldots; \hat{C}d_Z; \hat{C}t_1; \hat{C}t_2; \ldots; \hat{C}t_Z;] \tag{3.5}$$

where Z number of feature maps used.

3.4.4 External IR Feature

A good topic model learns common themes in the document collection. A limitation of our network is the lack of global- or collection-level information (as the input consists of only a document and topic). To incorporate collection-level information, we include an IR feature (F_{IR}) where we query document d_i using the topic words of t_i^j. We use Okapi BM25 (Robertson and Walker, 1994) to compute the relevance score of the document with respect to its N topic words independently, thereby constructing an N-dimensional feature vector.[2] This external feature vector is incorporated into the network after the convolutional layers (see Figure 3.1) and the vector P from Equation 3.5 is concatenated with F_{IR} to give P_{upd}

[2] $N = 5$ in our experiments.

$$P_{upd} = [\hat{C}d_1; \hat{C}d_2; \ldots; \hat{C}d_Z; \hat{C}t_1; \hat{C}t_2; \ldots; \hat{C}t_N Z; F_{IR}] \tag{3.6}$$

The document and topic hidden representations with the external IR vector are concatenated and fed to 2 dense layers and ultimately reduced to a sigmoid-activated score as given in Equation 3.7

$$\begin{aligned} H_1 &= f(W_{h1} \cdot P_{upd} + b_{h1}) \\ H_2 &= f(W_{h2} \cdot H_1 + b_{h2}) \\ O &= \sigma(W_o \cdot H_2 + b_o) \end{aligned} \tag{3.7}$$

where W_{h1}, W_{h2} and W_o are the weight matrices for the first and second dense layers, b_{h1}, b_{h2} and b_o are bias vectors, f is a non-linear activation function such as ReLU and σ denotes the sigmoid function.

3.4.5 Aggregating Human and System Scores for a Document

For each document we have a number of workers identifying the intruder topic. To aggregate the results, Chang et al. (2009) define model precision ($\texttt{mp}_{\text{GOLD}}$), which is the proportion of workers who correctly identified the intruder, as a proxy for how clearly the intruder topic is inappropriate for the document.

Our system and that of Bhatia et al. (2017) compute several scores for a document (one for each topic). Bhatia et al. (2017) select the topic with the maximum score as the intruder, and compute model precision (`mp`) based on that. This yields binary precision scores (i.e. the model either predicts the intruder correctly or not) and ignores the relative magnitude of the system score. We additionally propose using the normalised sigmoid score (`nss`) as a means of scoring the intruder topic for a given document, which is computed by normalising the raw sigmoid scores over all topics.

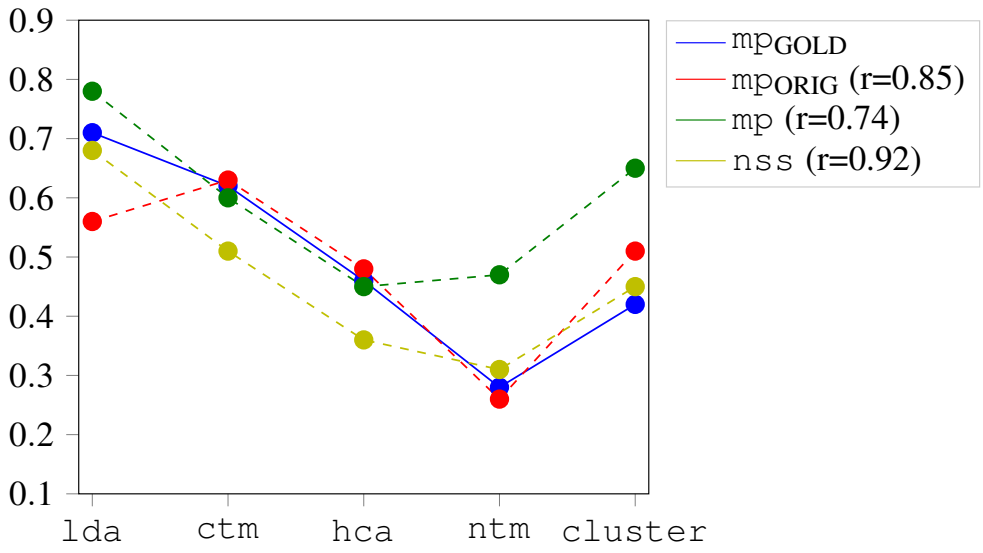


Figure 3.3: mp_{GOLD} vs. System Scores at the model level for APNEWS

3.4.6 Implementation Details

For our experiments, we train the model on outputs from all topics models over one dataset, and test it on the other (cross-domain training). We use a single channel for the convolutional networks, pad the documents as necessary ($k = 200$), and use the top-10 words to represent a topic (i.e. $m = 10$). Word embeddings are initialised using pre-trained GloVe (Pennington et al., 2014a) vectors ($d = 100$), and their weights are fixed during training. We use kernel windows of width $= \{3, 5, 7\}$ with 100 feature maps each and two (fully-connected) hidden layers, with dimensionality of 50 and 10. We use a dropout rate of 0.5, 0.5 and 0.25 after the document, topic and first hidden layer, respectively. We set the batch size to 100, and use Adam as the optimizer with a learning rate of 0.001. For activation functions, we use ReLU for the fully-connected layers and sigmoid for the final layer. To reduce variance, we run the models with 8 different seeds for initialisation and take the average for a topic's sigmoid score.

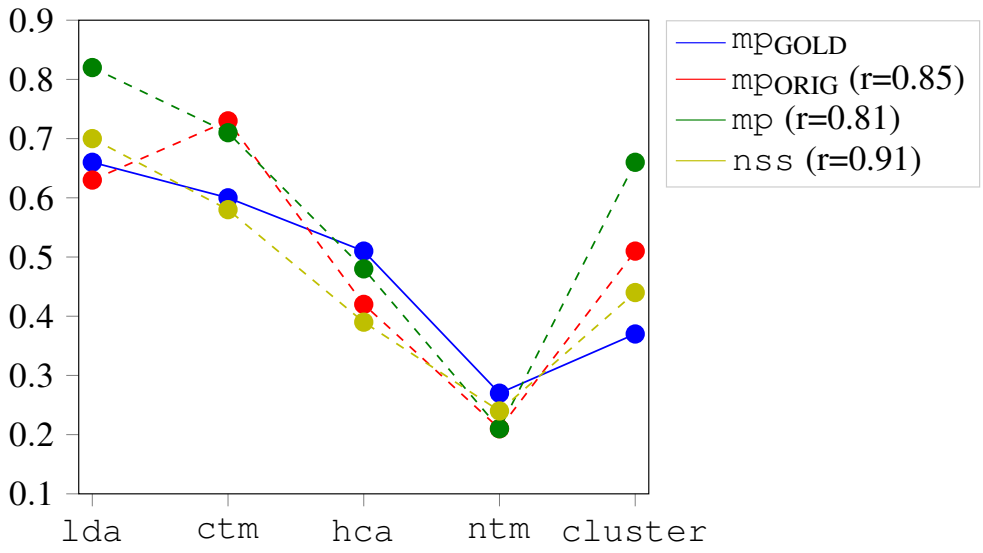


Figure 3.4: $\texttt{mp}_{\text{GOLD}}$ vs. System Scores at the model level for BNC

3.5 Results

By taking the mean of $\texttt{mp}_{\text{GOLD}}$ and mp over documents, Bhatia et al. (2017) compute a single human/system score for each topic model. Although this resulted in a strong correlation between $\texttt{mp}_{\text{GOLD}}$ and mp, the evaluation is limited to model-level comparison: it separates good topic models from bad topic models, but does not provide any insights into the performance of each top model over individual documents. We aim to improve model-level correlation in this work, in addition to analysing document-level evaluation, i.e. investigating how well the system predicts $\texttt{mp}_{\text{GOLD}}$ for each individual document.

We present plots of human and system scores for APNEWS and BNC in Figure 3.3 and Figure 3.4 respectively. There are 3 system scores: mp of Bhatia et al. (2017) ($\texttt{mp}_{\text{ORIG}}$), and mp and nss of our proposed system. In general, we found strong correlation for all systems, but nss of our proposed system performs substantially better than $\texttt{mp}_{\text{ORIG}}$, though our mp is lower than $\texttt{mp}_{\text{ORIG}}$.

To compare the performance of our system with human judgements at the document level, we compute mean absolute error (mae) between $\texttt{mp}_{\text{GOLD}}$ and nss/mp, as summarised in Table 3.1. We find for both datasets nss consistently outperforms $\texttt{mp}_{\text{ORIG}}$ and mp by a

Model	BNC → APNEWS			APNEWS → BNC		
	$\texttt{mp}_{\text{ORIG}}$	mp	nss	$\texttt{mp}_{\text{ORIG}}$	mp	nss
lda	0.47	0.31	**0.21**	0.40	0.32	**0.22**
ctm	0.44	0.34	**0.20**	0.41	0.31	**0.19**
hca	0.48	0.37	**0.21**	0.42	0.35	**0.20**
ntm	0.40	0.43	**0.19**	0.37	0.32	**0.18**
cluster	0.48	0.42	**0.19**	0.51	0.47	**0.22**
Overall	0.46	0.37	**0.20**	0.42	0.36	**0.21**

Table 3.1: mae between $\texttt{mp}_{\text{GOLD}}$ and nss/mp. "BNC → APNEWS" means the model is trained on BNC and tested on APNEWS. Boldface indicates optimal performance for each dataset.

Model	Best Topics	nss
lda	share revenue cents billion quarter earnings analysts net rose income	0.001
	european greece europe billion debt country crisis minister french france	0.002
ctm	building lodge bauer buildings fee part stephens hall property council	0.007
	military army afghanistan killed soldiers forces troops iraq war attacks	0.013
hca	shares earnings keywords insights profit thomson cents reuters premarket net	0.011
	upheld ruling appeals justices appellate supreme injunction plaintiffs unconstitutional rulings	0.051
ntm	rose shrank pct decliners quadrupled exhibitors parade spectrum index outperform	0.110
	arraigned burglarizing arrested bigamy detectives motorcyclist arraignment coroner accomplice fondled	0.141
cluster	soared plummeted climbed surged dipped tumbled dropped fell slipped rose	0.005
	students teachers kindergarten tutors elementary coursework curriculum teaching tutoring education	0.013

Table 3.2: Examples of best topics based on nss.

substantial margin, and also has a score close to human judgements. We can attribute this to the fact that nss provides more nuanced system predictions (over the full range $[0, 1]$), whereas mp tends to be binary.[3]

3.6 Discussion

One motivation we have in this work is to explore the use of topic intrusion as an alternative method for assessing topic models. Specifically, we aim to determine whether the best-

[3]Strictly speaking, it is continuous as it is averaged over the runs for the multiple random seeds, but in general, it tends to be (close to) 0 or 1.

identified topic models can still produce high-quality topics and if tnss can be effectively used to rank these topics. Given the encouraging mae results, we attempt to use nss to rank topics produced by a topic model.

To accomplish this, we first filter out the topics that occur in less than 5 documents as top 1-topic: these topics tend to be noisy, and as such do not appear with significant weight in any documents. For each of the filtered topics we randomly select 5—10 documents for which it is a top topic and calculate its mean nss over these documents. We then use the topics' mean nss to rank them; in Table 3.2 and Table 3.3 we show some selected best and worst topics for different topic models respectively. Overall, the top-ranked topics appear to be more descriptive than the bottom-ranked topics. Having said that, we found instances where coherent topics have low nss ranking (e.g. ctm topics in the bottom half of Table 3.2), but stress that ultimately the topic intrusion approach for assessing topics is very different to topic coherence.

As part of the topic intrusion, it is important to highlight the potential cases of uniform or constant probability mass. This means there could be three good topics for a document, all with uniform probability, or there could be cases where no good topics exist to allocate to a given document, once again meaning there will be a number of topics with middling probabilities (because they are all equally bad and there is no good topic, rather than they are equally good). The topic intrusion task remains effective in both scenarios. In the former case, where there are multiple good topics, humans will still be able to identify the intruder topic whereas in the latter case, where all topics are unsuitable, human would not be able to pick out the intruder topic, and so the topic intrusion task would be able to tell us that the former topic model's topics are good, but the latter topic model's topics are bad.

3.7 Conclusion

Topic coherence is traditionally used to rank/filter topics for end-user applications. However it tells us little about how well the topics describe the documents in the collection. In this chapter, we explored an alternative approach to evaluate topic models based on topic

Model	Worst Topics	nss
lda	lot good things long put start number making kind place	0.291
	political issue called issues policy decision long change statement support	0.271
ctm	online information internet book video media facebook phone computer technology	0.263
	show music film movie won festival tickets game band play	0.233
hca	richter riverboat sheppard lander plazas tam mandarin amarillo colosseum nassau	0.376
	deplorable interaction foresee envelope handwriting knot quickest scrambled alarmed mum	0.368
ntm	aboard spacewalks bushels budget lifeboats flotilla lifeboat spacewalk millage spaceflight	0.364
	evacuated evacuations evacuate evacuating airlifted twisters aftershocks evacuation driest barricaded	0.323
cluster	accord delegations accords cooperation consultations negotiators negotiation committees intergovernmental negotiations	0.323
	summaries summary critiques excerpts articles responses quotes references descriptions critique	0.309

Table 3.3: Examples of worst topics based on nss.

allocations in documents, i.e. via topic intrusion. We proposed an automated method that improves upon the state-of-the-art substantially at the model- and document-levels, and demonstrated that it can be used to rank/filter topics and plan to use it in subsequent chapters whenever topic models are employed. One of the tasks to explore for future would be intend to explore ways that can combine both the topic coherence and topic intrusion for topic model evaluation. Moreover, in this chapter we focussed on model document level evaluation, but topic models are also based on a collection and evaluation at the collection level is largely ignored and should be an area to explore for future work as well. We present some ideas around this in Chapter 7.

In this chapter, we explored an alternative method to optimize topics so that they accurately reflect the context of the document collection. In the subsequent chapters, we will dive deeper into CCS document collections and apply topic modelling - tuned using nss-methodology to see if it also aids in CCS detection.

Chapter 4

Climate Change Scepticism: Dataset and Detection

This chapter builds on:

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "You are right. I am ALARMED–But by Climate Change Counter Movement." arXiv preprint arXiv:2004.14907 (2020).

This chapter expands on the paper by going beyond its introduction of CCS literature from social science into the NLP domain and the presentation of a CCS dataset. While the paper did not provide a methodology for detecting CCS (and was solely focussed on bringing the problem to NLP domain and introduction of a dataset), this work introduces a novel approach based on pre-trained language models. It involves experimenting with different model configurations to effectively detect CCS, and also extending this methodology for shorter texts.

4.1 Introduction

Despite consensus in the scientific community about anthropogenic global warming, the web is awash with articles seeding with climate change scepticism (CCS). A major driver

of this content is opposing voices including the fossil fuel lobby, conservative thinktanks, big corporations, and digital/print media questioning and thus neutralising the science and research around climate change climate. We reviewed the literature about these organizations, their narrative, and language and examples in Section 2.2 and Section 2.3. More generally, we see that there is a formula consisting of a narrative structured around the principal ingredients of misleading information and propaganda, sprinkled with the stylistic elements of sensationalism, melodrama, clickbait, and satire. Table 4.1 throws light on one such article and other snippets of CCS articles were discussed before in chapter 2 in Table 2.1. Their approach broadly mirrors the strategies used in fake news in the political arena (Rashkin et al., 2017), but is specifically tailored to the domain of climate change. Thus the first point of tackling this is the detection of these articles, motivating the need of automatic detection or an alert system. In this chapter, we complement existing work on fact checking/fake news detection (Jiang and Wilson, 2018; Pérez-Rosas et al., 2018; Rashkin et al., 2017) by proposing the task of CCS detection.

We begin this chapter with a quick overview of language around CCS from Section 2.2 and pre-trained language models (PLMs) from Section 2.5. We then review the literature on language models scoring, particularly in the scoring of bidirectional language models through the means of pseudo log likelihood (PLL). To our knowledge there was no dataset in NLP for the domain of climate change scepticism. To be able to build a robust CCS detection system, the first step is to construct a dataset of articles which are CCS in nature. To this end, we scrape articles climate skeptic organisations (Section 2.2). Next, we introduce the CCS dataset and its corresponding carefully crafted test set from multiple sources. Following this, we introduce the task of CCS detection at a document level which is framed as a binary or a 1 class classification task.

In Chapter 3, we explored an alternative approach to topic model evaluation, which we experiment here to optimise our topic model parameters. Topic models have historically been the go to unsupervised approach in social sciences to study and draw insights from the literature, but as we will see in this chapter it may not be the right approach for the task of CCS detection, thus motivating the need to experiment with language models (LMs) and

'Climate Emergency' Fail of the Day – 25. As the media have collaborated to bombard the populace with environmental and climate doom propaganda, it is only fair and reasonable to hold their feet to the fire (as it were) and examine past media promoted predictions that have failed to materialise. Doubling down on the gone in eight years / gone in five years etc nonsense, the rhetoric was ramped up to gone in two years.... A new paper in the journal Nature argues that the release of a 50 Gigatonne (Gt) methane pulse from thawing Arctic permafrost could destabilise the climate system and trigger costs as high as the value of the entire world's GDP. The East Siberian Arctic Shelf's (ESAS) reservoir of methane gas hydrates could be released slowly over 50 years or "catastrophically fast" in a matter of decades – if not even one decade – the researchers said. Unfortunately, some real scientists did some research into this issue and produced a paper that showed the methane release was due to postglacial isostatic rebound rather than anthropogenic warming.

Table 4.1: An example of a CCS article

their perplexity score for our task. Generally speaking, perplexity is a measure of how well a language model can predict the likelihood of a sequence of words so lower the perplexity the better the language model predicts the next word. Thus, by continued training towards domain adaptation of a PLM on a CCS dataset — their perplexity score in an article should give us an idea on how much it reads article like a CCS document.

Following this, we conduct experiments across different models and settings i.e. PLMs with more parameters vs their smaller (distilled) versions, generic vs domain specific models, unidirectional LMs with casual language modelling objective vs bidirectional LMs with its masked language modelling objective. We extend the model to short texts and finally demonstrate that the method can be used to highlight spans of text which are potentially misleading.

4.2 Background

Public perception and reaction to climate change is a function of how the facts and narrative are presented to them (Fløttum, 2014; Fløttum et al., 2016), in large part because climate change is not just the physical science but has political, social, and ethical aspects (Flottum, 2017). In chapter 2, we reviewed the language around climate change and a range of topic

models and corpus linguistic methods that have been used to study the topical and stylistic aspects, including structured topic modelling (Roberts et al., 2014; Tvinnereim and Fløttum, 2015), keyphrase extraction via grammar induction (Salway et al., 2014), and analysis of frequently-used metaphors (Atanasova and Koteyko, 2017).

Although articles that express scepticism and misinformation are predominantly human-written, the recent emergence of large pre-trained language models means they can now be automatically generated. We discussed the literature around pre trained models in detail in Section 2.5.Radford et al. (2019) introduced a large auto-regressive model (GPT2) with the ability to generate high-quality synthetic text. One limitation of GPT2 is its inability to perform controlled generation for a specific domain, and Keskar et al. (2019) proposed a method to tackle this. Around the same time there also been a lot of success in bidirectional language models like BERT (Devlin et al., 2019) and its adaptations like ROBeRTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2019) and domain specific adaptations like ClimateBERT (Webersinke et al., 2021) with the distilled versions being highly efficient preserving upto 97 % of performance on downstream tasks with considerably less parameters. It is important to note at the time of this work these models were state of the art, but since then there has been development with the the likes of new state of the art models like GPT3 (Brown et al., 2020), ChatGPT [1] and GPT-4 (OpenAI, 2023) in the NLP ecosystem.

Bengio et al. (2003) introduced neural language modelling whereby we can infer the probability of a document by multiplying probabilities of each word in the context window given by:

$$\mathrm{P}(d) = \prod_{i=0}^{|d|} P(w_i|w_{<i}) \tag{4.1}$$

with its log version given by:

$$\mathrm{LP}(d) = \sum_{i=0}^{|d|} \log P(w_i|w_{<i}) \tag{4.2}$$

[1] https://openai.com/blog/chatgpt

This language modelling objective is used to estimate sentence probabilities for autoregressive models like unidirectional LSTMs (Hochreiter and Schmidhuber, 1997) or more recently models based on transformers architecture like GPT (Radford et al., 2018), GPT2 (Radford et al., 2019) and XLNET (Yang et al., 2019a). In terms of bidirectional models, pretrained transformers are largely based on masked language modelling objective in which a token is replaced with a *[MASK]* keyword and is predicted using bidirectional context. However the shortcoming of this bidirectional context is its inability to use Equation 4.1 to compute sentence probabilities, as it not based solely on left to right context. To address this, the approach employed is to mask words one at a time and calculate the probability of the masked word using bidirectional context (Lau et al., 2020; Shin et al., 2019)

$$\mathrm{P}(d) = \prod_{i=0}^{|d|} P(w_i|w_{<i}, w_{>i})) \tag{4.3}$$

with its log version given by and expressed in terms of pseudo log probability (PLP);

$$\mathrm{PLP}(d) = \sum_{i=0}^{|d|} \log P(w_i|w_{<i}, w_{>i}) \tag{4.4}$$

Strictly speaking, the sequence probability computed this way is not a *true probability value*, as these probabilities do not sum to 1.0 as is the case with Equation 4.1 over all sequences, hence the term *psuedo log likelihood*. But as Wang and Cho (2019) observe, there is no tractable means of computing true sequence probabilities for bidirectional models like BERT and similar models.

Diving deeper, Lau et al. (2020) studied sentence acceptability using both unidirectional and bidirectional models through a range of different metrics like Log Probability as in Equation 4.4, sentence length normalised versions like MeanLP and PenLP (Vaswani et al., 2017; Wu et al., 2016); unigram normalised versions like NormLP; and both unigram and length normalised versions together like SLOR (Pauls and Klein, 2012). They showed promising results for bidirectional models though at the same time also argued that the normalization

method employed also plays an important role for scoring sentence acceptability. In similar vein, Salazar et al. (2020) also explored masked language modelling scoring through PLP (Equation 4.4) and showed that bidirectional models demonstrate comparable performance to large unidirectional models like GPT2 for the tasks of Automated Speech Recognition and low resource neural machine translation.

4.3 Dataset

As a first step in building the pipeline for CCS detection we need to construct a new dataset. In this section we will dive into the scraping and construction of our dataset. We scrape articles with known climate change scepticism from 15 different climate sceptic organisations (Section 2.2. These organisations are selected from two sources: (1) McKie (2018); and (2) *desmogblog.com*,[2] a website that maintains a database of individuals and organisations that have been identified perpetuating climate change scepticism. We take the following into consideration when developing the dataset:

- We only scrape articles from organisations active in English-speaking countries: the United States, Canada, United Kingdom, Australia, and New Zealand.
- A considerable number of organisations are either dormant or have a very low level of activity. To make sure our dataset is up to date, we only scrape articles from organisations with a reasonable level of activity, i.e. they publish at least 1 article every month, and their latest publication is in 2020. [3]
- We set a minimum and maximum threshold of 10 and 400 articles respectively for each organisation. We set a maximum threshold so as to avoid bias towards one organisation. Note, however, that there is a considerable variance in the article length for different organisations. For instance, one organisation with only 10 articles has an average length of 342.1 words, while another organisation with 400 articles has an average length of 85.8 words.

[2] https://www.desmogblog.com/global-warming-denier-database
[3] This work for done in 2020 and hence we scrapped data upto that point in time.

# Organisations	12
# Articles	1168
Mean Length	559.3±640.2
Median Length	332.0

Table 4.2: Statistics of the training set

- As explained in Section 2.2, counter climate arguments can be broadly categorised into the science and policy frames. As organisations generally prefer one type of frame in their narrative, we manually identify frames associated with organisations, and select a set of organisations that produces a balanced representation of both frames in the dataset.

We split the documents into training and test partitions at the organisation level, where the training set comprises 12 organisations and the test set 3 organisations. We split at the organisation level because it allows us to test whether detection models are able to generalise their predictions to articles published by unseen/new climate change counter movement organisations. We present some statistics for the training documents in Table 4.2.

4.3.1 Test data

We extend our test set to include documents that do not have climate change scepticism, to create a standard evaluation dataset for detection. We collect documents from reputable sources, some of which are not climate-related, and some satirical in nature. Sources of the full test documents are as follows:

- **The Guardian**: A trusted source for independent journalism. We scrape articles under the category of climate change from both its U.K. and Australian editions. These articles test whether a detection system can correctly identify these articles as not containing climate misinformation.

- **BBC**: Similar to The Guardian, we scrape articles under the category of climate change.

Source	Snippet
Guardian	Plan to drain Congo peat bog for oil could release vast amount of carbon.The world's largest tropical peatlands could be destroyed if plans go ahead to drill for oil under the Congo basin, according to an investigation that suggests draining the area would release the same amount of carbon dioxide as Japan emits annually. Preserving the Congo's Cuvette Centrale peatlands, which are the size of England and store 30bn tonnes of carbon, is "absolutely essential" if there is any hope of meeting Paris climate agreement goals, scientists warn. However, this jungle is now the latest frontier for oil exploration .... Collaborations network that questions claims by developers that the oil deposit could contain 359m barrels of oil.....This untouched region is waterlogged for most of the year and is an important habitat for endangered forest elephants and lowland gorillas.
Beetota Advocate	PM Meets With Cricket Side To Discuss The 1.7m Hectares Of NSW Forests Destroyed By Bushfires. Not even six months after being officially elected as the Australian Prime Minister with absolutely no policies, let alone any acknowledgement of his government's denialism-led inaction on climate change, Scott Morrison has today had the opportunity to meet some more sportsmen! ... While the drought-stricken communities of rural Australian continue to burn at the hands of record-breaking and out-of-control bushfires, ScoMo has today met with the Australian cricket side for his ideal media appearance.. ..... The cricketers appeared distressed while also having to pose for goofy photos with the Prime Minister, ... planet's temperature that will result in the certain deaths of the billions of people that haven't been given permission to join Gina Rinehart and her Liberal Party employees in the spaceship.
Skeptical Science	Despite what you may read or see in the mainstream media, out in the real world, massive and rapid changes are taking place in many ecological systems as a result of global warming. The Earth seems to be already convinced of global warming and is responding quickly. Perhaps the most significant, and likely most enduring, are the shifts taking place in the Earth's oceans. Whilst many readers may have read or heard about Ocean Acidification, there are numerous other changes taking place in the oceans which should be equally as concerning. One such phenomena to appear in the last few decades is mass coral bleaching, a consequence of the continued warming of the oceans. ....Ocean Acidification in particular is a large looming threat (Veron 2009). The increasing frequency and severity of bleaching, coupled with the persistent decline in coral around the world, should however immediately dispel any myths about coral resilience.

Table 4.3: Snippet of articles from different sources

- **Newsroom**: This is a dataset released by Grusky et al. (2018) and consists of articles and summaries compiled from 38 different publications.[4] We take a random sample of articles which are not climate related (verified manually). These articles test whether a detection system is able to identify non-climate-related articles as not having climate misinformation.

- **Beetota Advocate**: This is a satirical website which publishes articles on current affairs happening locally and internationally;[5] we scrape articles related to climate change. Although there is a tone of sensationalism in the writing, the articles are created with the intent of humour. These articles test whether detection models are able to distinguish them from CCS articles, as both have similar stylistic characteristics.

- **Skeptical Science Arguments and Blogs (SS)**: This resource focuses on explaining what science says about climate change.[6] It publishes general climate blogs, and counters common climate myths by putting forth arguments backed by peer-reviewed research.

- **CCS**: These are articles from 3 CCS organisations, as detailed in Section 3.3. These documents are the only documents with climate change misinformation in the test data.

We present some examples of our test set, namely Guardian, Beetota Advocate, Skeptical Science and CCS, in Table 4.3 and previously in Table 4.1 respectively. We also present some statistics of the test set in Table 4.4.

4.4 Methodology

The detection task is a binary classification problem, where the goal is to classify whether a document is CCS in nature, i.e. if it espouses scepticism in climate change. Observing that CCS documents have unique stylistic (e.g. the use of exaggeration and sensationalism)

[4] https://summari.es/
[5] https://www.betootaadvocate.com/
[6] https://www.skepticalscience.com/

Type	Source	#Docs	Mean document length
	The Guardian	80	759.5±246.5
	BBC	60	696.2 ± 476.5
Non-CCS	SS	40	927.4± 510.3
	Newsroom	100	676.48± 625.2
	Beetota	60	360.7 ± 117.3
CCS	CCS articles	150	627.4 ± 301.9

Table 4.4: Test set document statistics.

and topical aspects (e.g. covering themes such as carbon tax and renewable energy), we explore the idea to further trained a pre trained language model (PLM) on the CCS training set (Section 4.3) - a strategy influenced by the approach by Gururangan et al. (2020). This method, essentially a form of domain adaptation, leverages the language model's inherent ability to capture both stylistic and topical patterns, thereby enhancing its effectiveness in performing the classification task. Henceforth we denote the positive class (CCS documents) as "$\texttt{ccs}^{+}$", and the negative class (non CCS documents) as "$\texttt{ccs}^{-}$".

To this end, we experiment using both autoregressive and bidirectional models. This includes autoregressive models like GPT-2 (Radford et al., 2019) and its distilled counterpart, DistilGPT2 (Sanh et al., 2019), as well as bidirectional models such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and the domain-adapted ClimateBERT (Webersinke et al., 2021), which was detailed in Chapter 2. In terms of model complexity, GPT-2 has approximately 124 million parameters, BERT has around 110 million, and the distilled models, including DistilBERT and DistilGPT2, feature about 66 million parameters each.

We further pretrain GPT2 and DistilGPT2 using its unsupervised language modelling loss (as described in Section 4.2), using the CCS training articles. After training, we compute the perplexity of a test document d as given in Equation 4.5. Perplexity is useful as is based on average log-likelihood of a document thereby normalising the length.

$$\begin{aligned} \mathrm{LP}(d) &= \sum_{i=0}^{|d|} \log P(w_i|w_{<i}) \\ \mathrm{PPL}(d) &= \exp\left(\frac{-\mathrm{LP}(d)}{|d|}\right) \end{aligned} \tag{4.5}$$

In GPT2, the model is trained with a maximum sequence length of B, i.e. self-attention is only performed for word tokens within a sequence that fits within this maximum length. When computing the probability of a word, $P(w_i|w_{<i})$, rather than considering all previous context words, we use only the previous $B-1$ words before w_i as context.

In case of the bidirectional models, we further pretrain using the masked language model on the CCS training articles, and the perplexity for a document d is computed in terms of pseudo log likelihood and its extension pseudo perplexity.

$$\begin{aligned} \mathrm{LP}'(d) &= \sum_{i=0}^{|d|} \log P(w_i|w_{<i}, w_{>i}) \\ \mathrm{PPL}'(d) &= \exp\left(\frac{-\mathrm{LP}'(d)}{|d|}\right) \end{aligned} \tag{4.6}$$

As mentioned in Section 4.2, sequence probabilities computed this way are not 'true probability values', but acts as a proxy for sentence evaluation. We use $(B-1)/2$ left and right context words to compute the word probability, and for words with less than $(B-1)/2$ left (right) context words, e.g. words appearing earlier (later) in the document, we use additional right (left) context words so that the total number of context words equals to $B-1$.

The perplexity of a document produced by these domain adapted language models tells us how much it reads like a CCS article: the lower the perplexity, the more likely. In preliminary experiments, however, we found that it was difficult to find a reasonable threshold to separate $\texttt{ccs}^+$ from $\texttt{ccs}^-$ content. Given that, we instead measure the "perplexity difference" between the adapted and original off the shelf pre-trained models.

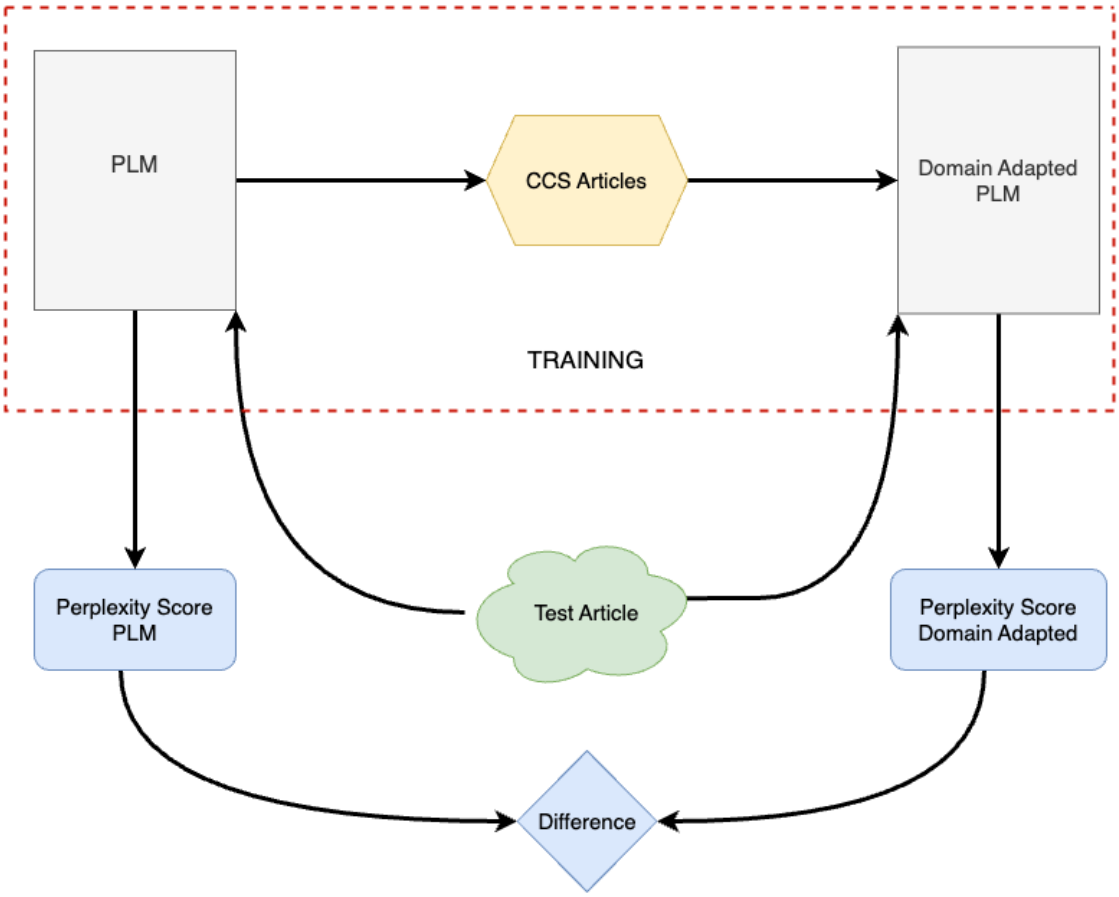


Figure 4.1: Architecture diagram for Methodology

$$\begin{aligned} \triangle\text{PPL}(d) &= \text{PPL}_a(d) - \text{PPL}_o(d) \\ \triangle\text{PPL}'(d) &= \text{PPL}'_a(d) - \text{PPL}'_o(d) \end{aligned} \tag{4.7}$$

where PPL_a and PPL_o refer to the document perplexity produced by the domain adapted and original/off-the-shelf autoregressive models like GPT2 and DistilGPT2. Similarly, PPL'_a and PPL'_o refer to the document perplexity produced by the fine-tuned and original/off-the-shelf bidirectional models like BERT, DistilBERT and ClimateBERT. The intuition behind this formulation is simple: for a CCS document, PPL_a (PPL'_a) should be low but PPL_o (PPL'_o) should be high ($\triangle$PPL $\downarrow$); and for a $\texttt{ccs}^-$ document we should have the reverse ($\triangle$PPL $\uparrow$).

figrefppl-ch4 gives an overview of the basic methodology. We introduce a threshold based on the perplexity difference to classify whether a document is $\texttt{ccs}^{+}$ or $\texttt{ccs}^{-}$.

4.5 Experiments

In this section, we present the results of our experiments for different models. We also include the topic models from Chapter 3 and a 1-class Support vector machine (SVM) baseline. We employ SVM as before deep learning approaches became the norm, SVM were the de-facto models for text classification. The following list shows all the models used in our experiments.

1. 1-class SVM model (Schölkopf et al., 2000). A traditional SVM works by separating two classes using a hyperplane with the largest possible margin whereas a 1-class SVM is primarily employed for novelty or anomaly detection where they learn a decision boundary of the feature space that contains the majority of the data, and anything outside that region is considered an anomaly. They work by finding a hyperplane that maximally separates the training data from the origin in the feature space. We use TF-IDF weighted unigram and bigrams as our features in the feature space. We call this model `1-SVM`.

2. LDA based topic models whose hyper-parameters of number of topics (n), a prior to control topic-document distribution (α), and a prior to control topic-word distribution (β) were optimised using 2 different evaluation metrics; (1) topic coherence with the help of "CV" which is based on co-occurrence and the score is calculated with the help of Normalised pointwise mutual information and cosine similarity (Röder et al., 2015) and; (2) using the normalised sigmoid score (NSS) introduced in the last chapter. We train the topic model using the train docs and post training calculate perplexity of the test docs and tune a threshold value on the development set which acts as a boundary for classification of CCS articles. Henceforth, we call topic model optimized with 'CV' and 'NSS' as `TM-CV` and `TM-NSS` respectively.

3. fine-tuned BERT binary classifier (Devlin et al., 2019). As this is a binary classifier we need to prepare training data with both $\texttt{ccs}^+$ and $\texttt{ccs}^-$ instances. We use the train CCS articles (Section 4.3) for $\texttt{ccs}^+$ and randomly sample equal number of documents from newsroom (Grusky et al., 2018) for $\texttt{ccs}^-$, where half are in the domain of climate change.[7] We call this model as `2-BERT` and note that this is a supervised model.

4. Domain adapted GPT-2 (see Section 5.4) model, where we further pretrain using the train partition of CCS articles: `U-GPT`

5. Domain adapted DistilGPT2 model: `U-DGPT`

6. Domain adapted BERT language model: `U-BERT`

7. Domain adapted DistilBERT model: `U-DBERT`

8. Original ClimateBERT model further pretrained on CCS articles.[8] The idea is to check if the ClimateBERT which has some topical knowledge is able to adapt to CCS articles: `U-CBERT`

9. We use the original ClimateBERT model alongside a domain-adapted DistilBERT (`U-DBERT`), as each is designed to target different aspects of the dataset. Typically, as outlined in Equation 4.7, we calculate the $\triangle$ between a domain-adapted model and its original or off-the-shelf counterpart. However, in this case, we compare the domain-adapted model used on CCS articles (`U-DBERT`) with ClimateBERT (which serves as the off-the-shelf model here). It's crucial to note that while we do not further train the off-the-shelf ClimateBERT, it is a domain-adapted model trained on general climate change text. Thus, our hypothesis is that ClimateBERT will primarily focus on non-CCS articles, whereas the `U-DBERT` is expected to target CCS articles hence theoretically $\triangle$PPL should enable us to distinguish between $\texttt{ccs}^+$ and $\texttt{ccs}^-$ articles.

[7] Based on the presence of one of the two phrases: *global warming* and *climate change*.

[8] Important to note that in a sense ClimateBERT itself is a domain adapted on climate change related datasets

Topic Words
cost emission policy economic tax high government carbon energy economy
agenda policy sceptic united-nations activist conference society climate international
poor poverty fraud green climate cost reduce co2 emissions truth
electricity build wind-turbine battery storage solar-panel turbine power storage

Table 4.5: Topics from LDA optimised with NSS

We revisit and further elaborate on the intuition behind this approach in Section 4.6. This model is referred to as U-CDBERT.

We give implementation details later in Section 4.8.

4.6 Results and Discussion

We present Precision, Recall and F-Score results over the $\texttt{ccs}^{+}$ class in Table 4.6. Looking at the topic models, we see that they perform poorly although TM-NSS does slightly better than TM-CV. This is not a surprise given that topic models are more useful in terms of coarse grained topic exploration, but they might struggle to capture language styles due to its bag of words input. We present a few topics from this model in Table 4.5 which are quite coherent e.g. they cover concepts related to "carbon tax" (first topic), "organizations and activists" (second and third topic) and "renewables" (last topic).

Next moving over to the language models, we first focus on the autoregressive models of U-GPT and U-DGPT. We see that both U-GPT and U-DGPT perform consistently well in terms of both Precision and Recall. Interestingly, U-DGPT despite being a smaller model than U-GPT performs at par (a few points better) than U-GPT. Moving to bidirectional models we can make 3 observations: (1) the overall performance of bidirectional models (like U-BERT, U-DBERT, U-CBERT and U-CDBERT) is a bit lower than autoregressive models; (2) U-CBERT despite of being a climate domain specific model, performs the worst on overall F-Score even though it has the best Precision; and (3) U-CDBERT which uses standard CBERT with a fine tuned DBERT does the best of all bidirectional models. In

Section 4.2 we reviewed prior studies that found bidirectional models work better as language modelling scores (Lau et al., 2020; Salazar et al., 2020) but we do not have the same findings for our task. We hypothesise this could be due to: (1) It is ever less clear what the pseudo perplexity difference ($\triangle$PPL$'$) means when we are working with pseudo log likelihoods; and (2) the nature of our task is different to sequence to sequence neural machine translation or automatic speech recognition tasks.

Turning next to the binary classifier `2-BERT` we see it has the best `Recall` but suffers from low `Precision`, resulting in an overall poor `F-Score`; for an alert system this means it will raise alarms for many documents that do not actually have climate scepticism (which will impact on user experience). `1-SVM`, which is a simple baseline, unsurprisingly also results in poor performance.

Model	Precision	Recall	F-Score
U-GPT	0.71	0.75	0.73
U-DGPT	0.73	0.77	0.75
U-BERT	0.71	0.67	0.70
U-DBERT	0.71	0.68	0.70
U-CBERT	0.73	0.53	0.62
U-CDBERT	0.68	0.72	0.71
2-BERT	0.40	0.95	0.55
TM-CV	0.34	0.68	0.45
TM-NSS	0.36	0.70	0.46
1-SVM	0.35	0.71	0.46

Table 4.6: Classification performance ($\texttt{ccs}^+$ class).

To understand better the discrepancies between `Precision` and `Recall` for the various language models of `U-GPT`, `U-DGPT`, `U-DBERT`, `U-CBERT` and `U-CDBERT`, we present a breakdown of predictions over different sources in the test set in Table 4.7 and Table 4.8. To reiterate the first 5 sources (The Guardian to Skeptical Science) are documents without any scepticism. From this, we can see that in general the unidirectional language models have a balanced performance, classifying most articles correctly, but also being more conservative and missing more CCS articles. Comparing `U-GPT` and `U-DGPT`, we see that

Dataset (#Docs)	U-DGPT		U-DGPT	
	ccs$^+$	ccs$^-$	ccs$^+$	ccs$^-$
The Guardian (80)	17	63	11	69
BBC (60)	0	60	3	57
SS (40)	2	38	15	25
Newsroom (100)	3	97	3	97
Beetota (60)	27	33	10	50
CCS (150)	112	38	115	35

Table 4.7: Predictions of U-GPT and U-DGPT over different test sources

Dataset (#Docs)	U-DBERT		U-CBERT		U-CDBERT		2-BERT	
	ccs$^+$	ccs$^-$	ccs$^+$	ccs$^-$	ccs$^+$	ccs$^-$	ccs$^+$	ccs$^-$
The Guardian (80)	19	71	18	62	10	70	74	6
BBC (60)	1	59	18	42	13	47	42	18
Skeptical Science (40)	9	31	35	5	11	29	40	0
Newsroom (100)	6	94	7	93	6	94	10	90
Beetota (60)	14	46	25	35	8	52	35	25
CCS (150)	106	44	110	40	103	47	144	6

Table 4.8: Predictions of U-DBERT, CBERT, U-CDBERT and 2-BERT over different test sources

U-GPT works better for Skeptical Science source whereas the performance of U-DGPT over Beetota is better

Focussing on bidirectional models we see that all 3 models achieve a comparable performance for the CCS class with U-CBERT (domain adapted ClimateBERT on CCS articles) doing the best. An important observation is that while U-DBERT does well over ccs$^-$ sources, U-CBERT struggles especially over the BBC, Beetota and skeptical science sources. This issue might stem from the fact that ClimateBERT was already domain adapted on a large climate change corpus and after additional training on CCS articles (resulting in U-CBERT), it appears to be overfitted to the CCS domain, leading to difficulties in distinguishing between CCS and non-CCS content. This observation underscores the potential benefits of experimenting with a combined model called as U-CDBERT. This approach would maintain the original climateBERT and use it in with the domain-adapted U-DBERT to calculate $\triangle \text{PPL}'$

At Climate Conference Angus Taylor has been treated to a rousing reception at an international Climate Change conference today. As huge parts of his home country burn or smoulder as a result of catastrophic bushfires, which experts and scientists say is exacerbated by climate change, the Australian Energy and Emissions Reduction Minister has won the crowd over by discrediting the threat of the very thing the entire international conference is on. Speaking at the Conference of the Parties to the United Nations Convention on Climate Change, Angus Taylor someone how managed to prevent the rest of the group from laughing at him for his nation's inaction on Climate Change. Seated amongst representatives of small Pacific Islands whose lives are threatened by rising sea levels, Angus Taylor managed to sway the crowd on the hot button issue by hitting out at 'scaremongering lefties pushing a radical agenda.'..... "I went there expecting to be laughed out of the house because they all believe in climate change, can see our nation is on fire and are aware I'm not taking it seriously," Taylor said over the phone. "But they lapped it up."

Table 4.9: Example of Beetota Article wrongly classified by all models

with the aim to leverage the strengths of both models. In the results shown in Table 4.8, U-CDBERT demonstrates improved performance across all $\texttt{ccs}^-$ classes with a particularly strong performance for Betoota correctly classifying 52 out of 60 instances as $\texttt{ccs}^-$.

Moving to 2-BERT, we can make three observations: (1) the majority of climate-related documents from The Guardian, BBC and Skeptical Science are classified as $\texttt{ccs}^+$, indicating a topical bias (i.e. it tends to classify climate-related documents as $\texttt{ccs}^+$); (2) due to the satirical and sensationalism nature of Beetota articles, most articles (35 out of 60) are classified as $\texttt{ccs}^+$; and (3) the performance over Newsroom is very strong. We attribute the last observation to the presence of Newsroom articles in the training data, demonstrating the brittleness of supervised models.

It is also important to point out that some of the wrong classifications may be less attributed to the quality of model or training but to the difficulty of the task and test set, especially in the case of Beetota as its language occasionally mirrors CCS. We share one such example in Table 4.9 which classified as $\texttt{ccs}^+$ by all the models. Looking at the example we can see that it has phrases like "scaremongering lefties pushing a radical agenda" which is inline with the language we see in CCS articles, giving it a very low $\triangle$PPL ($\triangle\text{PPL}'$ in

case of bidirectional models) value thus confusing the models and driving the classification for the whole article towards $\texttt{ccs}^{+}$.

Document Text
Inside the Cult of Climate Supremacy, this month has seen a little flurry of pop-psychology op-eds appearing in the news by a coterie of Climate Supremacists masquerading as academics. These blinkered, dare I say, hooded darlings of the hard-left have no interest in informing; they're only interested in denigrating those that don't genuflect to the altar of man-made global-warming ... Describing skeptics directly as "people who are less educated" the poor possum had trouble concentrating, it seems, while putting together his puerile piffle; ... But if you believe Mr Hall's lack of self-awareness is remarkable, just wait for the next instalment on the personalities inside the cult of Climate Supremacism
Are Climate Models Overpredicting Global Warming? Many recent climate models have been predicting dire global changes. The problem is climate forecasters currently ignore decades of scientific best-practices that would offer more accurate predictions. Thankfully, there are attempts to rectify the truly dodgy methodology that has been used to crank out forecasts of 21st-century climate ... The authors of the new paper show that the aggregate models are making huge errors in three of the places on earth that are critical to our understanding of climate ... It's high time that the scientific community come clean about longstanding climate shenanigans. Averaging up a large number of models that don't work well is guaranteed to produce an unreliable forecast.
Take a Look at the New Consensus on Global Warming. A scientific consensus has emerged among top mainstream climate scientists that "skeptics" or "lukewarmers" were not long ago derided for suggesting — there was a nearly two-decade long "hiatus" in global warming that climate models failed to accurately predict or replicate ... More importantly, the paper discusses the failure of climate models to predict or replicate the "slowdown" in early 21st century global temperatures ... Democrats and environmentalists praised Karl's work which came before the Obama administration unveiled its carbon dioxide regulations for power plants ... Then, in early 2016, mainstream scientists admitted the climate model trends did not match observations – a coup for scientists like Patrick Michaels and Chop Knappenberger who have been pointing out flaws in model predictions for years.

Table 4.10: Detected climate scepticism spans from `U-GPT`. Red refers to spans with lowest $\triangle$PPL ; Orange still highlighting scepticism but not as high as red and black being non sceptic content. Lower perplexity means higher likelihood to have scepticism.

4.7 Span Highlighting

As an alert system for flagging climate change scepticism in web articles, it would be useful to further highlight parts of the content that expresses these scepticism sentiments. The advantage of using a language model for the detection task is that we can look at individual token probability distributions and use it to detect spans of text that carry this scepticism. For a predicted $\texttt{ccs}^+$ article, we can perform sentence segmentation to get a set of sentences and compute $\triangle$PPL (Equation 4.5) over sentences to extract sentences with the lowest perplexity difference.[9] We present an example of detected spans for several documents in Table 4.10.

For the first document, we can see it is written with emotionally-charged words (a typical writing style in CCS articles), and as such all sentences have very low $\triangle$PPL. The overall perplexity difference of the document is also very low (-25.2), suggesting that the `U-GPT` is confident that it is a $\texttt{ccs}^+$ document. For the second and third document, the style is more subtle, although the red and orange highlighted sentences are climate-sceptic statements that question the science of climate models. These results suggest that span highlighting can be a useful feature in a climate misinformation alert system to reveal areas in the document that users should pay more attention to.

4.8 Implementation Details

As `2-BERT` is a binary classifier, we needed to prepare training data with both $\texttt{ccs}^+$ and $\texttt{ccs}^-$ instances as mentioned in Section 4.5. We randomly sampled 10% of articles from this data to use as a development set and used the rest as training articles. To this end, we used uncased BERT-Base as the pre-trained model and finetuned for classification over the *[CLS]* token, and the following hyper-parameters: batch size $= 6$, number of epochs $= 2$, and gradient accumulation $= 3$. We implemented `2-BERT` using the SimpleTransformers library. [10]

[9] The probability of each word is calculated the same as before using $B-1$ context words. The difference here is that we are normalising over sentences, rather than the whole document.

[10] https://github.com/ThilinaRajapakse/simpletransformers

For U−GPT, we use the small size GPT-2 (Radford et al., 2019) and the following configuration: $B = 512$, batch size $= 2$, learning rate $=$ 5e-5 (default), max gradient norm $=$1.0 (default), gradient accumulation $=$2 and number of epochs $= 3$. Similarly for U−BERT we use uncased BERT-Base with similar configuration with the masking rate $= 0.15$. We implemented the models using the transformers library[11].

For topic modelling we used mallet-lda [12] and gensim. [13]. The models were optimised using tthe wo coherence metrics: CV and NSS. This optimization process determined the optimal number of topics n to be 32 for CV and 36 for NSS. Additionally, the optimal α values were 0.03 for CV and 0.02 for NSS. For both models, the β value was set at 0.01. For 1−SVM, we used TF-IDF weighted unigram and bigrams and implement the model using sklearn.[14]

4.9 What about Short Texts?

In this section, we explore if the detection models mentioned above can be applied to short texts without further training. To this end we collect $\texttt{ccs}^+$ short texts from two sources:

1. misleading or inaccurate climate claims as flagged by *climatefeedback.org*,[15] which is a fact checking website that verifies the credibility of climate-related claims. This source has texts which are topically similar to $\texttt{ccs}^+$ texts.

2. a small collection of tweets by Donald Trump (45th President of the United States of America).[16] This source has writing style of sensationalism and exaggeration as also seen in $\texttt{ccs}^+$ articles.

In case of $\texttt{ccs}^-$ short texts, our dataset comprises the following:

[11] https://github.com/huggingface/transformers
[12] https://mimno.github.io/Mallet/
[13] https://radimrehurek.com/gensim/models/ldamodel.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html
[15] https://climatefeedback.org/claim-reviews/
[16] We would like to point out that since scrapping a few of his tweets, his Twitter account had been banned and only reinstated back in November 2022

Type	Source	#Docs
ccs$^-$	Tweets	45
	SS + NASA	55
	Beetota	40
	Newsroom	40
ccs$^+$	climatefeedback.org	60
	Trump Tweets	20

Table 4.11: Short text test set statistics.

1. Climate change arguments in Skeptical Science and NASA;[17]

2. A collection of climate change related tweets by Bernie Sanders, UN Climate Change, António Guterres (Former Secretary General of UN), Carolina Schmidt (President of COP25), and Patricia Espinosa C (Executive Secretary of UN Climate Change)

3. Headlines from Beetota Advocate and Newsroom as mentioned in Section 4.3.1 but instead of whole articles just the headlines.

These collected ccs$^+$ and ccs$^-$ texts constitute the short text test set. We present the statistics in Table 4.11 and a few examples in Table 4.12. For consistency and to be able to directly compare short text performance with that of long text, we make sure that the ratio of ccs$^+$ and ccs$^-$ instances in the short text test set is similar to that for long text. We use the models we trained previously (U-GPT, U-DGPT, U-DBERT and U-CDBERT) and apply them directly to these short texts.

We present the ccs$^+$ performance for short texts in Table 4.13. Overall, we still see that autoregressive models U-GPT and U-DGPT have similar performance, with U-GPT doing slightly better than the latter. Both of them have the best performance relative to other models, although results drop compared to the long text performance. Precision is noticeably lower here, indicating an increase in false positives. We present a sample of false positives in Table 4.15. The first example bears the sensationalist style of climate CCS articles, and due to a lack of context, is classified as ccs$^+$. The second example is factually correct but still

[17] https://climate.nasa.gov/scientific-consensus/

Source	Example
climatefeedback.org	Scientists were caught 'adjusting' sea level data to create false impression of rising oceans .
Trump Tweets	For those that constantly say that "global warming" is now "climate change"—they changed the name. The name global warming wasn't working.
ccs$^-$ Tweets	Today, we need to reduce emissions by 7.6 each year. So, it is imperative that governments not only honour their national contributions under the Paris Agreement, they need to substantially increase their ambition.
SS + NASA	Between 1950-2009 sea level at Tuvalu rose at the rate of 5.1 (±0.7) mm per year. This is almost 3 times larger than average global sea level rise over the same period.
Beetota	Barnaby Says Lets Not Get Political Because Half A Billion Dead Animals Probably Voted Greens

Table 4.12: Examples of short text

Model	Precision	Recall	F-Score
U-GPT	0.47	0.78	0.59
U-DGPT	0.47	0.75	0.58
U-DBERT	0.41	0.71	0.52
U-CDBERT	0.46	062	0.53

Table 4.13: Classification performance (ccs$^+$ class) for short texts.

gets wrongly detected as ccs$^+$. We suspect this is because in quite a few CCS articles, there is a tendency to quote climate facts in order build their counter argument, and as a result the sentence may appear like a typical sentence in a CCS article.

Table 4.14 provides a breakdown of U-GPT, U-DGPT, U-DBERT and U-CDBERT performance for different sources. We see U-GPT and U-DGPT yield a very similar trend with small variations in Newsroom and Beetota again highlighting the fact that the smaller model are able to perform on par with the bigger variant in this task. For both U-GPT and U-DGPT, most of the false positives seem to come from Skeptical Science/NASA and Beetota. In the case of U-DBERT we see that even though the performance on classifying CCS sources like climatefeedback and trump tweets is comparable to U-GPT and U-DGPT, the false positive rate is very high for tweets, highlighting the fact that the model is failing to differentiate between general climate related tweets and the CCS short texts. Moving on to U-CDBERT, it achieves better Precision and is able to get the source tweets right (we suspect due to the involvement of climate BERT it could identify the the trustworthy tweets)

Dataset (#Docs)	U-GPT		U-DGPT		U-DBERT		U-CDBERT	
	ccs$^+$	ccs$^-$	ccs$^+$	ccs$^-$	ccs$^+$	ccs$^-$	ccs$^+$	ccs$^-$
Tweets (45)	18	27	21	24	36	9	4	41
SS+NASA (55)	21	34	21	34	17	38	27	28
Beetota (40)	26	14	22	18	22	18	18	22
Newsroom (40)	6	34	3	37	7	33	10	20
climatefeedback.org (60)	46	14	44	16	41	19	35	25
Trump Tweets (20)	16	4	16	4	15	5	14	6

Table 4.14: Predictions of U-DBERT and U-CDBERT over different test sources (correct predictions in bold).

Examples
You are damn right I am alarmed. Climate change is a major national security threat and a global emergency..
While summer maximums have showed little trend, the annual average Arctic temperature has risen sharply in recent decades.

Table 4.15: Examples of short text misclassified as ccs$^+$ by U-GPT.

but ended up misclassifying the *climatefeedback* source a lot more, resulting in much lower recall.

4.10 Conclusion

We introduced the task of climate change scepticism detection, and developed a dataset made up of documents written by counter climate change movement organisations, balanced up against documents from a range of non-CCS sources. We then proposed novel models and showed them to be effective for both document classification and CCS span detection. We experimented with topic models and though got granular concepts out but that did not help us in distinguishing between CCS and non-CCS sources. In terms of language models unidirectional models did better over our task than bidirectional models. This needs more exploration e.g. in the direction of development of new metrics for bidirectional models for this task similar to the work done by Lau et al. (2020). Although there were some

encouraging results for applying our method to short text, there is still work to be done more to explore in the area of short text climate scepticism classification. We believe either further pretraining a language model on CCS short texts (or joint training with long texts) should help.

In Section 2.3 we reviewed the concept of framing and how arguments on climate skepticism are categorized into Science and Policy. We believe that incorporating these signals into our classification system, either through a multi-task objective or by adding features, could enhance performance. Additionally, it is worth noting that our dataset comprises documents from various genres and writing styles, including reports, comments, and pseudo-scientific reports and adding some meta features highlighting those differences could also help in improving the classification performance

So far we have discussed detection of climate sceptic articles and identification of highly misleading spans in the CCS articles. In the next chapter, we will dive the identification of more fine grained classes of arguments used in the writing of such texts.

Chapter 5

Automatic Classification of Neutralization Techniques in the Narrative of Climate Change Scepticism

This chapter builds on the paper:

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "Automatic classification of neutralization techniques in the narrative of climate change scepticism" In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2021.

In this chapter we add (other than what is already in the paper) a background section on semi-supervised learning, provide a more detailed explanation of our methodology and annotation process, and introduce new experiments that integrate topic models with a pre-trained BERT model.

5.1 Introduction

The previous chapter explored the task of detecting articles and text spans in them that exhibit climate change scepticism (CCS). While that acts as valuable first step in better understanding climate change scepticism, there is a need to better understand the narrative of an article more deeply e.g. how it frames an argument.

Public perception is influenced by the narrative presented to them (Fløttum, 2014; Fløttum et al., 2016), and one of the tools CCS texts employ to build counter-climate narratives (McKie, 2018) is "neutralization". We reviewed the literature around the theory of neutralization and framing in Section 2.3 but briefly it is defined as justification/vindication for a deviant behaviour (Kaptein and Van Helvoort, 2019; Maruna and Copes, 2005; Sykes and Matza, 1957). Thus here we propose a method to automatically classify these neutralization techniques (henceforth "NT"), as a tool to analyse CCS narrative at scale and help build counter-narratives. Table 5.1 presents two examples of neutralization in the context of climate change; more examples were presented previously in Chapter 2.

To be able to build a model to classify NT, we need labelled data. In situations where annotated data is limited, alternative approaches need to be explored such as semi-supervised learning. We commence this chapter by reviewing the literature around semi-supervised learning, ranging from self training, multi-view training, variational auto encoders and data augmentation methods to more sophisticated methods that combine multiple techniques (Chen et al., 2020). Next, we point out a few applications of semi-supervised learning in other social science domains, thus providing the motivation for us to explore this approach paradigm for our task in this chapter.

Following this, we revisit the 7 neutralization techniques and 2 frames which were discussed previously in Chapter 2. As multiple neutralization techniques may be used together to construct a CCS argument, we frame our task as a multilabel classification problem. We next describe how we construct our data and annotation procedure. We then experiment with different model configurations i.e. fine tuned `BERT`, domain-adapted `BERT`, a topic enhanced `BERT` where we optimise the number of topics using the methodology in Chapter 3, and add inferred topic distributions as supplementary features, and `BERT`

Sure, we should reduce greenhouse gases, but if our climate policies hurt our ability to create more wealth and bring power to the world's poor, then we are ridding the patient of the disease, but only by killing him
It's very convenient for alarmist greens to blame the fires of Australia and California on global warming. In reality, global warming is just a natural cycle and the policies they themselves advocate are the culprits.

Table 5.1: Neutralization examples

configurations with semi-supervised training objectives. We show that our best model performs on par with humans performance and end the chapter by also revisiting the CCS spans detected in Chapter 4 (Table 4.10) and employing neutralization on those spans to give a nuanced understanding of those sceptic spans.

5.2 Background

In Chapter 2 we gave a detailed overview on general misinformation and will give a brief recap here. Research on fake news and propaganda has primarily operated at the article level, and focused on binary detection (presence vs. absence) (Barrón-Cedeno et al., 2019; Rashkin et al., 2017). Da San Martino et al. (2019) argued for the need for finer granularity in propaganda detection, both in terms of propaganda sub-types and fragment-level detection. In a similar vein, Nakamura et al. (2020) proposed fine-grained classes of fake news to differentiate between misleading, manipulated, or totally false content. More recently in the climate change domain, Luo et al. (2020) released a stance-annotated dataset for global warming, and proposed an opinion framing task to study discourse used in the debate around global warming.

One challenge in building supervised NLP models is the strong dependency on labelled data. To tackle this, one approach is to apply transfer learning from pretrained language models i.e. using large general pretrained models and fine tune it on task-specific labelled data as discussed in Section **??** (Conneau and Lample, 2019; Devlin et al., 2019; Peters et al., 2018b; Radford et al., 2019; Yang et al., 2019a). Another approach is semi-supervised

learning with self training (Rosenberg et al., 2005; Triguero et al., 2015; Yarowsky, 1995) being one of the traditional approach. It acts as a teacher-student framework where a supervised classifier first learns from labelled examples and then this classifier is used to make predictions on unlabelled examples. The labelled set is augmented with the most confident predictions on the unlabelled set and the classifier is re-trained; the process is repeated in an iterative manner.

Another approach widely employed in semi-supervised learning is co-training (Balcan et al., 2004; Blum et al., 2004; Wang and Zhang, 2006) where separate classifiers are trained on distinct subsets of the labelled data, and the most confident predictions are added to the labelled data set of the other classifier. Building on this, Sindhwani et al. (2005); Sindhwani and Rosenberg (2008) proposed co-regularization by sharing a single objective function between different supervised classifiers. Together these methods are also referred to as multi-view training, as they give multiple different 'views' of the data (Van Engelen and Hoos, 2020). Taking inspiration from multi-view training, Clark et al. (2018) introduced a cross-view training framework, which again works as as a teacher-student framework where the teacher teaches the student to make predictions on unlabelled data but by adding auxiliary prediction modules that see only a limited or restricted view of the data.

More recently, variational auto encoders (VAEs) (Kingma and Welling, 2013) have been leveraged in a semi-supervised setting to reconstruct sentences and predict the labels with the help of a latent variable and a distribution parametrised by μ and σ. But variational inference struggles with instability, and training collapses in a textual setting due to the vanishing Kullback–Leibler (KL) divergence loss term and needs tricks for optimization (Bowman et al., 2015). Yang et al. (2017) showed encouraging results with a VAE in a sequence to sequence setting with a regular `LSTM` as an encoder and dilated convolutions (Van den Oord et al., 2016) as a decoder. Building on it, Gururangan et al. (2019) proposed a system which employed `VAE` with ideas from pre-training language models. They pretrained a deep VAE on unlabelled text, then concatenated it (with its internal layers frozen) with task specific features for downstream labelling tasks.

Elsewhere, techniques like data augmentation and interpolation have been employed to increase the size of the training data. Data augmentation has been widely applied in computer vision by applying transformations either through their geometry - flipping or rotating them at different angles, or color mixing by adding or removing different shades of hue (Baird, 2007; Perez and Wang, 2017; Shorten and Khoshgoftaar, 2019). Extending the idea to textual data, Wei and Zou (2019) experimented with synonym replacement whereas Kumar et al. (2019) presented with novel paraphrase formulation through the means of submodular function maximization to induce diversity in paraphrases to augment textual data. Xie et al. (2020a) used back translations, i.e. translating to a pivot language (like German or Russian) and translating it back to English (Sennrich et al., 2016a), for consistency regularization to produce paraphrases on unlabelled data. Similarly, Zhang et al. (2017) proposed a method called 'Mixup', which operates by creating virtual training samples through linear interpolations of labeled data points primarily used to interpolate images. It creates new data points by merging two images and their labels (Berthelot et al., 2019; Verma et al., 2019) i.e. given two labeled data points (x, y) and ($x^{'}$, $y^{'}$), where $\hat{x}$ represents an image and $\hat{y}$ is the one-hot representation of the label, Mixup creates virtual training samples as follows:

$$\begin{aligned} \hat{x} &= \text{mix}(x, x^{'}) = \lambda x + (1-\lambda)x^{'} \\ \hat{y} &= \text{mix}(y, y^{'}) = \lambda y + (1-\lambda)y^{'} \end{aligned} \tag{5.1}$$

where $\lambda \in [0,1]$.

More recently, pretrained models and semi-supervised learning have been combined with much success, e.g. Xie et al. (2020a) used BERT along with consistency regularization on unlabeled data and Croce et al. (2020) extended the fine-tuning process of `BERT` to a generative adversarial setting. Drawing from various strategies and combining them, Chen et al. (2020) devised a data augmentation approach of Tmix and its semi-supervised variant MixText (`MTEXT`). However, applying Mixup directly to text data is challenging due to

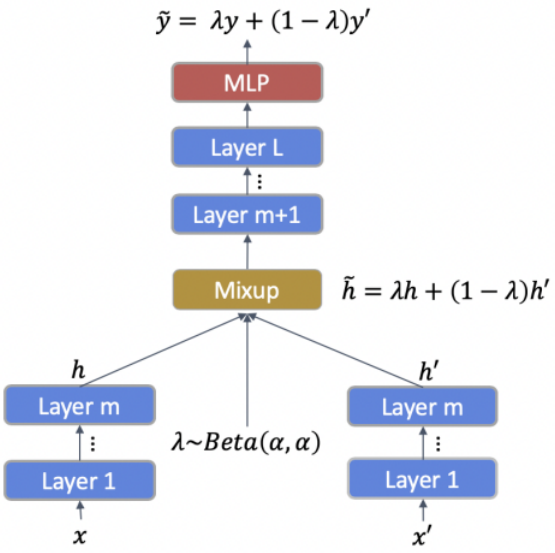


Figure 5.1: Architecture of `TMix`, reproduced from Chen et al. (2020)

the discrete nature of text tokens. Tmix addresses this issue by interpolating within the textual hidden space, mixing up the hidden representations of two training instances (selected from both labeled and unlabeled data) within BERT. This process creates a large number of enhanced data samples, with the linear interpolation guided by a mixture parameter λ sampled from a Beta distribution. For an encoder with L layers, Tmix chooses to mix up the hidden representation at the m-th layer, $m \in [0, L]$ as shown in the architecture diagram of Tmix presented in Figure 5.1. `MTEXT`, on the other hand, is a semi-supervised learning framework that utilizes `TMix` as a data augmentation approach. We present the architecture of `MTEXT` in Figure 5.2. The key idea behind `MTEXT` is to leverage unlabelled (X_u) data in addition to the labelled data (X_l) for training the model. `MTEXT` uses back translation based data augmentation (K augmentations) on unlabelled data, with label assignment based on pseudo labels for unlabelled and augmented data (X_k) and the final inference a weighted average in a teacher -student self training manner, to generate diverse paraphrases while preserving the semantics of the original sentences. After inferring the labels for unlabelled instance, a superset is constructed which is the concatenation of the labelled set, unlabelled set and unlabelled augmented set (X_l, X_u, X_k) thereby increasing the total available training samples.

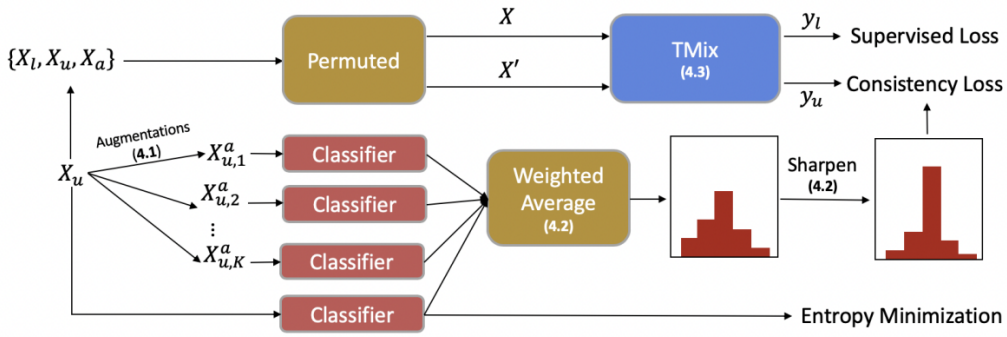


Figure 5.2: Architecture of MTEXT, reproduced from Chen et al. (2020)

During training, 2 data points are randomly sampled from this superset where the process of *Tmix* is repeated and KL-divergence used as a loss function. In the case of both samples coming from the labelled set this turns into a supervised objective expressed through standard cross-entropy loss which quantifies the dissimilarity between the predicted label distribution and the actual labels. The framework also incorporates an entropy minimization (Grandvalet and Bengio, 2005) loss function that encourages the model to assign sharp probabilities to unlabelled data samples i.e. the model is incentivized to assign high probabilities to one class and low probabilities to the others. This helps to boost performance, especially when dealing with scenarios involving a large number of classes.

Semi-supervised learning has found a wide array of applications, especially in the social sciences, due to its effectiveness with small data sizes. One such application is in the study of speech, initially introduced by Austin (1975) suggests that language serves not only to describe the world but also to perform actions and distinguishes between locutionary acts (the act of saying something), illocutionary acts (the intention behind the statement), and perlocutionary acts (the effect on the listener). Building on this Searle (1976) further developed the theory by categorizing illocutionary acts into specific types such as assertives, directives, commissives, expressives, and declarations therby providing a more structured framework for understanding. Speech Act Theory also involves analyzing various communicative acts, ranging from asking questions and making comments to offering suggestions. These acts can be domain-specific, including commissive actions like directives, assertions, expressions, pledges, or sentence specificity (Li and Nenkova, 2015; Subramanian et al., 2019a). Zhang

et al. (2012) employed transductive SVM (Joachims et al., 1999) with graph-based label propagation (Zhou et al., 2003; Zhu et al., 2003) to label unlabelled data, using only a small seed labelled training set for speech act recognition. Similarly, Joty and Mohiuddin (2018) modelled speech acts for asynchronous conversations through a domain adversarial learning approach in a semi-supervised setting. Following this, Subramanian et al. (2019b) introduced target-based speech act classification to study political campaigns, leveraging in-domain unlabelled data with cross-view semi-supervised training (Clark et al., 2018) in conjunction with contextualized embeddings and metadata. In a deeper exploration of election discourse and political manifestos, Subramanian et al. (2019a) performed fine-grained pledge specificity prediction (7-levels of specificity, (Pomper and Lederman, 1980)) on various policy issues by modelling it as an ordinal regression task, employing cross-view training with adjustments for ordinal regression and real-valued outputs. Ito et al. (2019) used semi-supervised models to improve decision making on medical data. They employ generative adversarial networks (GANs) (Goodfellow et al., 2020) to generate artificial data targeting imbalanced/minority classes, and then combine self training with co-training to improve the confidence of predictions on the unlabelled set thereby increasing the labelled set at every iteration. In this chapter, we will take the work on semi-supervised applications forward and apply it to the domain of climate change scepticism.

We reviewed the literature around neutralization and framing in Section 2.3 but for continuity we give a brief recap here. Dunlap and Brulle (2015), Farrell (2016), and Boussalis and Coan (2016) classified arguments related to CCS into two frames: science ("SCIENCE") and policy ("POLICY"). The SCIENCE frame challenges the scientific evidence and may include denial or promotion of pseudoscience, whereas the POLICY frame focuses on cost and economy-related issues such as carbon tax, targeting scientists or shifting the responsibility of action to other countries. McKie (2018) modified Sykes and Matza (1957)'s original NT schema and renamed the CCS arguments related to neutralization by linking it with the SCIENCE and POLICY frames to produce the 7 NT classes: Denial of Responsibility (Deny-Responsibility $\rightsquigarrow$SCIENCE), Denial of Injury1 (Deny-Injury1 $\rightsquigarrow$SCIENCE), Denial of Injury2 (Deny-Injury2 $\rightsquigarrow$SCIENCE), De-

Argument or Example	**NT**	**Frame**
There's no indication this is anything other than natural variability, with humans not playing a part	Deny-Responsibility	SCIENCE
There is a very real probability that global warming has been overestimated by computer models, and won't be too bad	Deny-Injury1	SCIENCE
CO2 is plant food and good for the planet, as it is essential for plants in photosynthesis	Deny-Injury2	SCIENCE
Despite forecasts of warming, the world has actually been cooling, so global warming is a hoax	Deny-Victim	SCIENCE
An avalanche of global warming alarmism is about to hit, thanks to environmentalists, the media, and a few scientists	Condemn	POLICY
So-called "new renewable energy technologies" are extremely expensive and rely on huge subsidies, pushing up energy costs	Loyalties	POLICY
New Zealand's actions should be less ambitious than Australia's because Australia is a wealthier country	Justify	POLICY

Table 5.2: Examples of counter climate arguments and their frames reproduced from Table 2.4

nial of Victim (Deny-Victim $\rightsquigarrow$SCIENCE), Condemnation of the Condemner (Condemn $\rightsquigarrow$POLICY), Appeal to Higher Loyalties (Loyalties $\rightsquigarrow$POLICY), Justification by Comparison (Justify $\rightsquigarrow$POLICY). We gave examples of these 7 neutralization techniques in Table 5.2, reproduced from Table 2.4. CCS texts often use multiple neutralization strategies together. as seen in the second example in Table 5.1 where Condemn (POLICY) is used to blame the *alarmist greens* and Deny-Responsibility (SCIENCE) is used to highlight that global warming is a *natural cycle*.

5.3 Dataset

We introduce our dataset in this section. To be able to train our models, we need labelled examples based on the neutralisation classes. To this end, we construct our NT dataset from 3 sources:

1. We extract paragraphs from CCS documents we scraped across 15 Climate Change counter movement organizations, comprising the training set from the previous chapter (Section 4.3).

2. We compile a set of CCS sentences and paragraphs from the work of McKie (2018).[1]

3. We use anti-global warming opinions (sentences) from the work of Luo et al. (2020). We used opinions/stances which disagree with the statement: *climate change/global warming is a serious concern*.

This results in a mixture of sentences and paragraphs, resulting in diversity in the dataset (with longer snippets expected to have more multi-labelling). We henceforth call these text snippets "sentences" for brevity. Our dataset has a total of 8000 sentences, of which 785 were annotated (and the remainder used as unlabelled data).

5.3.1 Annotation and Mechanical Turk

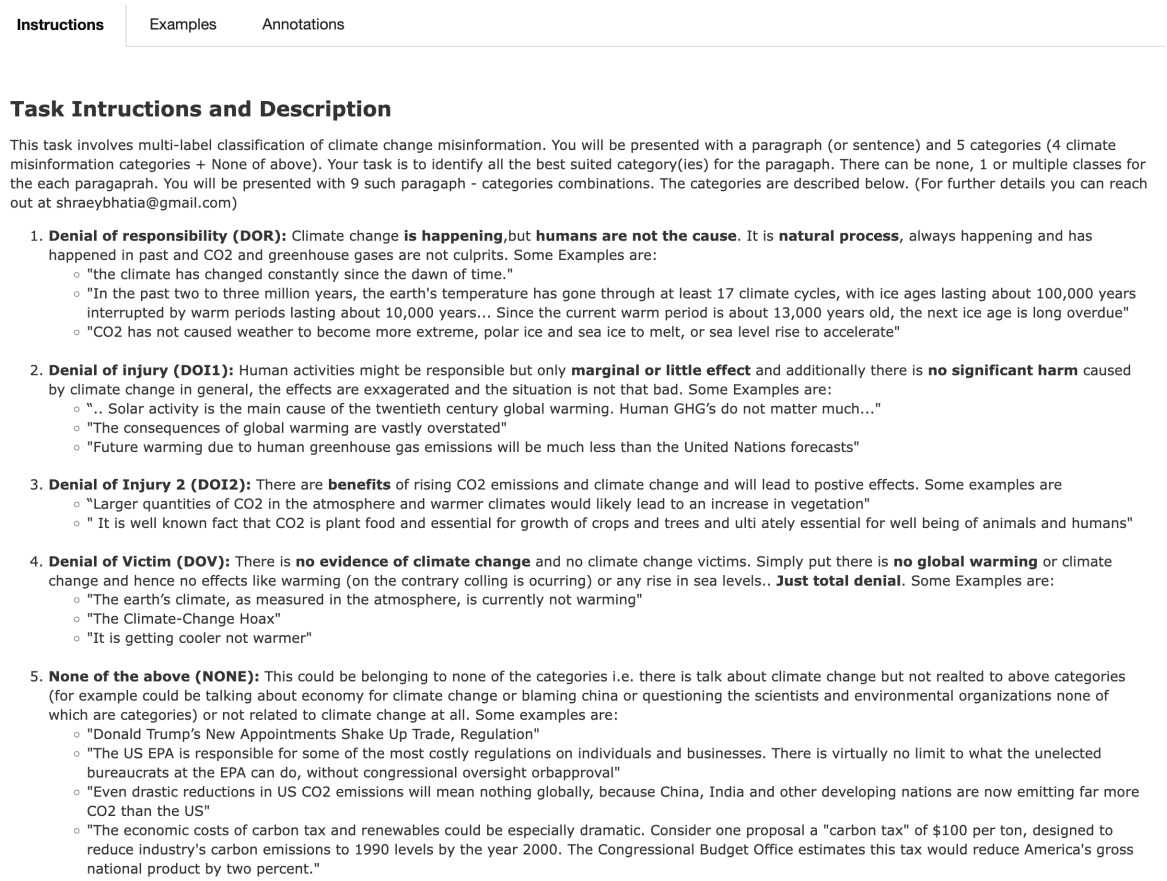
In this section, we describe the design and mechanics of our crowd sourcing experiments to label the spans with NT and frame labels. For the model to be able to categorise NT strategies and their frames correctly we need high quality annotated data. We formulate the task as a multi-label classification problem where an annotator selects `NONE`, or one or more NT labels. To make the task easier for annotators, we split it into 2 annotation subtasks based on the two frames:

- the `SCIENCE` frame, which consists of `Deny-Responsibility`, `Deny-Injury1`, `Deny-Injury2`, and `Deny-Victim`
- the `POLICY` frame, which consists of `Condemn`, `Loyalties`, and `Justify`

i.e. we have annotators label examples to the `SCIENCE` and `POLICY` frames separately, to simplify the annotation process.

We combine annotations by taking a majority vote for each frame, and label a sentence as `NONE` only if it is the majority-class for both sub-tasks i.e. none of the NT labels are majority-assigned in either frame. We collect human judgements using Amazon Mechanical Turk with 9 sentences forming a single HIT, one of which acts as a quality control in the

[1] Extracted from the appendix of their thesis



Instructions | Examples | Annotations

Task Intructions and Description

This task involves multi-label classification of climate change misinformation. You will be presented with a paragraph (or sentence) and 5 categories (4 climate misinformation categories + None of above). Your task is to identify all the best suited category(ies) for the paragaph. There can be none, 1 or multiple classes for the each paragaprah. You will be presented with 9 such paragaph - categories combinations. The categories are described below. (For further details you can reach out at shraeybhatia@gmail.com)

1. **Denial of responsibility (DOR):** Climate change **is happening**,but **humans are not the cause**. It is **natural process**, always happening and has happened in past and CO2 and greenhouse gases are not culprits. Some Examples are:
   - "the climate has changed constantly since the dawn of time."
   - "In the past two to three million years, the earth's temperature has gone through at least 17 climate cycles, with ice ages lasting about 100,000 years interrupted by warm periods lasting about 10,000 years... Since the current warm period is about 13,000 years old, the next ice age is long overdue"
   - "CO2 has not caused weather to become more extreme, polar ice and sea ice to melt, or sea level rise to accelerate"

2. **Denial of injury (DOI1):** Human activities might be responsible but only **marginal or little effect** and additionally there is **no significant harm** caused by climate change in general, the effects are exxagerated and the situation is not that bad. Some Examples are:
   - ".. Solar activity is the main cause of the twentieth century global warming. Human GHG's do not matter much..."
   - "The consequences of global warming are vastly overstated"
   - "Future warming due to human greenhouse gas emissions will be much less than the United Nations forecasts"

3. **Denial of Injury 2 (DOI2):** There are **benefits** of rising CO2 emissions and climate change and will lead to postive effects. Some examples are
   - "Larger quantities of CO2 in the atmosphere and warmer climates would likely lead to an increase in vegetation"
   - " It is well known fact that CO2 is plant food and essential for growth of crops and trees and ulti ately essential for well being of animals and humans"

4. **Denial of Victim (DOV):** There is **no evidence of climate change** and no climate change victims. Simply put there is **no global warming** or climate change and hence no effects like warming (on the contrary colling is ocurring) or any rise in sea levels.. **Just total denial**. Some Examples are:
   - "The earth's climate, as measured in the atmosphere, is currently not warming"
   - "The Climate-Change Hoax"
   - "It is getting cooler not warmer"

5. **None of the above (NONE):** This could be belonging to none of the categories i.e. there is talk about climate change but not realted to above categories (for example could be talking about economy for climate change or blaming china or questioning the scientists and environmental organizations none of which are categories) or not related to climate change at all. Some examples are:
   - "Donald Trump's New Appointments Shake Up Trade, Regulation"
   - "The US EPA is responsible for some of the most costly regulations on individuals and businesses. There is virtually no limit to what the unelected bureaucrats at the EPA can do, without congressional oversight orbapproval"
   - "Even drastic reductions in US CO2 emissions will mean nothing globally, because China, India and other developing nations are now emitting far more CO2 than the US"
   - "The economic costs of carbon tax and renewables could be especially dramatic. Consider one proposal a "carbon tax" of $100 per ton, designed to reduce industry's carbon emissions to 1990 levels by the year 2000. The Congressional Budget Office estimates this tax would reduce America's gross national product by two percent."

Figure 5.3: A screenshot of the annotation guidelines for the SCIENCE frame

form of a data instance from McKie (2018). Each HIT was annotated by a minimum of 5 and maximum of 10 annotators.[2]

The annotation task for the SCIENCE frame is shown in Figure 5.3 (annotation guidelines), Figure 5.4 (annotation examples) and Figure 5.5 (annotation interface). Though the figures only show the annotation task for SCIENCE frame, a similar process was used for the POLICY frame thus giving us annotations for all the 7 classes. The design of the task is divided into 3 tabs; the first tab of *Instructions* (annotation guidelines) describes the task in detail, explaining the different NT classes. The second tab of *Examples* provides 3 examples, and their correct labels with the required explanation and rationale behind them, and the last

[2]The crowdsourcing experiments were conducted before the university required ethical approval for Mechanical Turk Tasks and stricter minimum pay requirements came into effect.

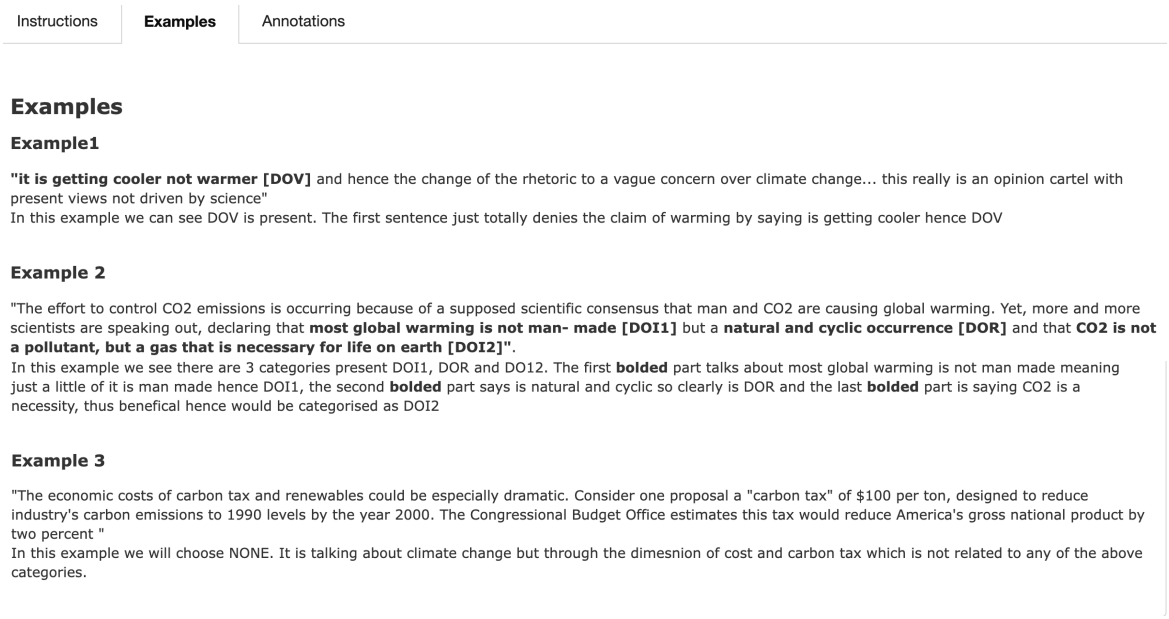


Figure 5.4: Labelling examples for the SCIENCE frame

tab of *Annotation* is the actual annotation interface. The design layout allows annotator to switch between different tabs during the task.

To pass quality control for a given HIT, the annotator has to select the correct class for the quality control sentence (which is not flagged in any way to the annotator, and presented in random order); the annotations from a worker are discarded if their average pass rate across all HITs attempted is ≤ 0.7. We collect additional annotations by releasing the task internally to a small number of local workers. We restrict the task to workers with an approval of 97%+, based in the US, Canada, UK, Australia, or New Zealand.[3] Each HIT was paid at USD $0.61, and took an average of 5 minutes to complete. This amounts to $7.32 per hour, which is slightly above the US federal minimum wage ($7.25) at the time the task was carried out.[4]

We present statistics of the labelled data in Table 5.3. Interestingly, we see 3 large classes of NT — Deny-Victim, Condemn, and Loyalties— implying that most CCS narratives completely deny climate change, condemn the scientists, or prioritise the economy.

[3]It is important to point out that due to the polarization of the climate change issue, there will be an element of political and geographical bias for climate change related task

[4]The US federal minimum wage could have changed since then

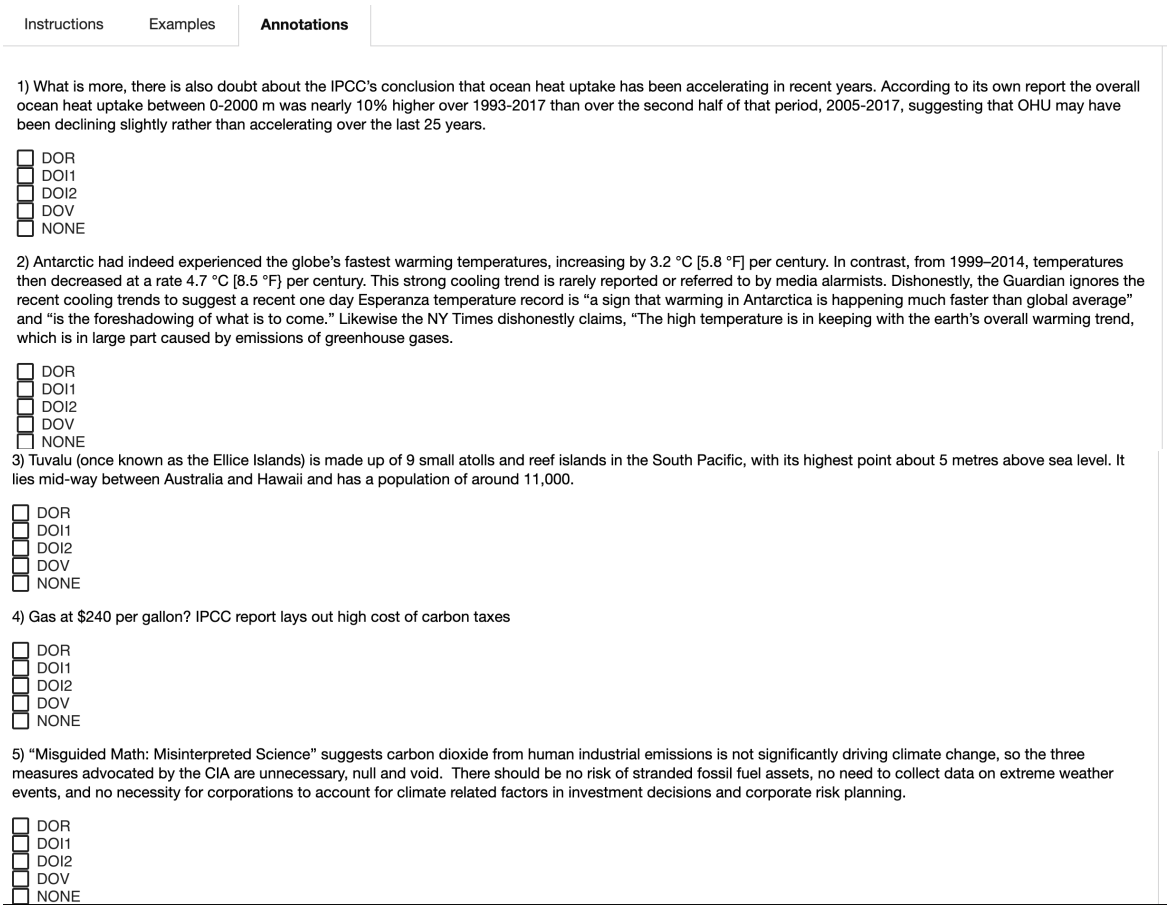


Figure 5.5: The annotation interface for the SCIENCE frame

5.4 Automatic Classification

In this section, we present the task in more detail, and the different model architectures for the classifiers in use. We formulate the task as a multi-label classification with the goal of detecting the NT classes employed in the sentence. We experiment with an SVM as a baseline and then explore several BERT-based supervised and semi-supervised models. As it is a multilabel classification problem, we add a number of one-vs-rest classification layers (one for each class) on top of BERT, and update all parameters during fine-tuning. In addition, following Gururangan et al. (2020), we also experiment with adaptive pretraining for BERT, i.e. before we fine-tune BERT to our task, we pretrain the off-the-shelf BERT using the masked language model (MLM) objective on the CCS document set we described in Chapter 4.

NT	%	Sentence Length
Deny-Responsibility	11.47	44.43
Deny-Injury1	8.78	44.71
Deny-Injury2	9.67	41.56
Deny-Victim	23.18	42.31
Condemn	35.67	49.89
Loyalties	21.23	48.87
Justify	4.01	50.09
NONE	7.52	36.31

Table 5.3: Distribution across classes.

1. **SVM**: we adopted a standard linear-kernel SVM in one vs. rest mode to a multilabel setting (Schölkopf et al., 2000) where each class is treated as an independent binary classification problem (one class is positive and the rest are negative).

2. **BERT**: Standard supervised BERT (Devlin et al., 2019) fine-tuned using the labelled data.

3. **BERT$_{\texttt{topic}}$** Peinelt et al. (2020) proposed a topic informed BERT which combines topic model features with BERT for the task of semantic similarity detection and proved particularly effective in domain-specific scenarios. Taking inspiration from this, we also experiment with a similar methodology by first training LDA (Blei et al., 2003) topic model on the collection of CCS documents. These documents were gathered from 15 different Climate Sceptic organizations as detailed in Chapter 4. Once the LDA topic model is trained, we move on to the inference phase. Here, we determine the topic distribution for the input sentence and combine it with the [CLS] token from BERT. The parameters of the topic model are optimised using the Normalised Sigmoid Score (NSS), a metric we introduced in Chapter 3.

4. **MTEXT**: A semi-supervised BERT-based model based on Chen et al. (2020) detailed in Section 5.2 but extended to a multilabel setting. The supervised objective ($\mathscr{L}_s$) uses standard cross-entropy loss whereas the unsupervised objective uses consistency loss ($\mathscr{L}_{cl}$) in the form of KL-divergence, and an entropy minimization loss $\mathscr{L}_{em}$ for the

unlabelled set is added, yielding the overall objective $\mathscr{L}_{nt} = w_1\mathscr{L}_s + w_2\mathscr{L}_{cl} + w_3\mathscr{L}_{em}$, where w_x are tunable hyper-parameters.

5. **MTEXT$_{\texttt{multi}}$**: As we saw in Section 5.2, NT is associated with SCIENCE and POLICY frames. We experiment with adding these frames (including the NONE class, 3 in total) as an auxiliary objective, creating another supervised loss ($\mathscr{L}_{frame}$). $\mathscr{L}_{frame}$ is implemented as multilabel loss, as a sentence can have both SCIENCE and POLICY frames. The final objective is $\mathscr{L}_{nt} + \alpha\mathscr{L}_{frame}$, where α is a tunable hyper-parameter.

6. **BERT***: Following Gururangan et al. (2020), we also experiment with adaptive pretraining for BERT, i.e. before we fine-tune BERT to our task, we pretrain the off-the-shelf BERT using the masked language model objective on CCS documents from previous chapter. Models with adaptive pretraining are marked with '*'

7. **BERT$^*_{\texttt{topic}}$**: Similar to above, BERT$_{\texttt{topic}}$ but with adaptive pretraining.

8. **MTEXT***: MTEXT with adaptive pretraining.

9. **MTEXT$^*_{\texttt{multi}}$** MTEXT$_{\texttt{multi}}$ with adaptive pretraining.

At test time, we add two extra post-processing rules for the NONE class: (1) it is automatically selected if all other classes are predicted to be absent; and (2) it is never selected if any other classes are predicted to be present.

We split the labelled data into train/dev/test with 450/135/200 sentences. The semi-supervised models (MTEXT variants) also have access to the unlabelled 7215 sentences. We use the uncased BERT-base as the pretrained model for all experiments. We detail the full training details and hyper-parameters in Section 5.4.1.

We present Precision, Recall and F-Score results for the test-set in Table 5.4. To provide an approximate upper bound, we also present the estimated human performance, which is computed by randomly isolating a worker's annotations, and calculating the agreement with the rest, repeating this 100 times to reduce variance, and averaging the results.

Model	Precision	Recall	F-Score
BERT	0.57	0.62	0.59
BERT*	0.60	0.64	0.62
$\text{BERT}_{\text{topic}}$	0.63	0.56	0.59
$\text{BERT}^{*}_{\text{topic}}$	0.63	0.54	0.58
MTEXT	0.62	0.71	0.66
MTEXT*	0.63	0.71	0.67
$\text{MTEXT}_{\text{multi}}$	**0.64**	**0.73**	**0.68**
$\text{MTEXT}^{*}_{\text{multi}}$	0.62	0.71	0.67
SVM	0.78	0.39	0.49
Human	0.69	0.72	0.70

Table 5.4: NT multi-label classification performance.

We first look at the (fully) supervised results, and see that the baseline BERT performs the worst, but adaptive pretraining (BERT*) boosts performance. Looking at $\text{BERT}_{\text{topic}}$ we see it is marginally better than BERT but loses its competitive with adaptive pretraining i.e F-Score of BERT* > $\text{BERT}^{*}_{\text{topic}}$.

Moving on to semi-supervised models (MTEXT, MTEXT*, $\text{MTEXT}_{\text{multi}}$ and $\text{MTEXT}^{*}_{\text{multi}}$), we see consistent gains, highlighting the benefits of using unlabelled data. $\text{MTEXT}_{\text{multi}}$ with its multi-task objective gives a small but appreciable gain over MTEXT, producing performance that is on par with human performance. Interestingly though, adaptive pretraining (MTEXT* and $\text{MTEXT}^{*}_{\text{multi}}$) does not seem to be of much help. We suspect this is because both techniques are broadly based on the similar idea, i.e. to improve performance by leveraging additional unlabelled data. In MTEXT* we use adaptive pretraining for the model to adapt to the climate sceptic data. In the case of $\text{MTEXT}^{*}_{\text{multi}}$ we try to achieve a similar objective using unlabelled examples but with a different technique. The baseline SVM has the lowest performance

To better understand how "data efficient" these models are, we present F-Score over varying amounts of labelled training data for BERT, BERT*, MTEXT and $\text{MTEXT}_{\text{multi}}$ in Figure 5.6. We see that MTEXT and $\text{MTEXT}_{\text{multi}}$ outperform BERT and BERT* substantially

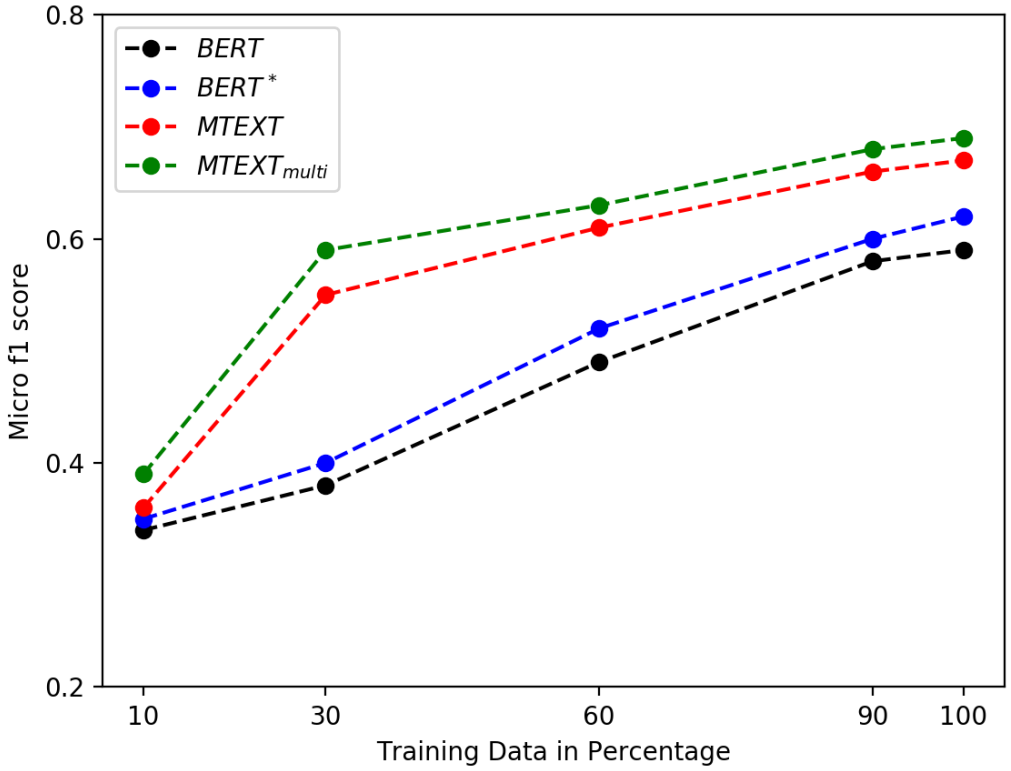


Figure 5.6: `F-Score` performance over increasing amounts of training data.

with only 30% of training data (135 instances), and maintain their strong performance as data quantity increases, showing that in situations with few labelled examples semi-supervised learning is even more beneficial.

Finally, we present a breakdown of F1 scores for each class in Table 5.5. Adaptive learning mostly improves the two large classes (`Deny-Responsibility` and `Loyalties`) for `BERT` vs. `BERT`*. When we incorporate semi-supervised learning (`MTEXT` and $\texttt{MTEXT}_{\texttt{multi}}$), we see large improvements for all the small classes (`Deny-Injury1`, `Deny-Injury2`, and `Justify`), suggesting that semi-supervised learning benefits the smaller classes most. To get an estimated upper bound we also present human F1 scores for each class. Looking at those scores, we observe that the gap with human performance is higher for the smaller classes (`Deny-Injury1`, `Deny-Injury2`, and `Justify`) even for our best model, highlighting the limitations of semi-supervised learning.

Model	**Deny-Responsibility**	Deny-Injury1	Deny-Injury2	**Deny-Victim**
BERT	0.51	0.13	0.52	0.62
BERT*	0.57	0.13	0.49	0.64
MTEXT	0.68	0.40	0.73	0.65
$\text{MTEXT}_{\text{multi}}$	0.62	0.50	0.73	0.70
SVM	0.30	0.08	0.56	0.38
Human	0.64	0.62	0.88	0.65
Model	**Condemn**	**Loyalties**	Justify	NONE
BERT	0.72	0.72	0.00	0.20
BERT*	0.73	0.80	0.00	0.20
MTEXT	0.73	0.76	0.30	0.30
$\text{MTEXT}_{\text{multi}}$	0.73	0.80	0.35	0.30
SVM	0.68	0.67	0.00	0.00
Human	0.77	0.81	0.61	0.56

Table 5.5: F1 breakdown across classes in SCIENCE and POLICY frame. The largest classes are bolded.

Span	**NT Prediction(s)**
The authors of the new paper show that the aggregate models are making huge errors in three of the places on earth that are critical to our understanding of climate. It's high time that the scientific community come clean about longstanding climate shenanigans.	Condemn
A scientific consensus has emerged among top mainstream climate scientists that "skeptics" or "lukewarmers" were not long ago derided for suggesting — there was a nearly two-decade long "hiatus" in global warming that climate models failed to accurately predict or replicate	Deny-Injury1 Condemn

Table 5.6: Examples of NT predictions on CCS Spans

In the previous chapter, we proposed a model for CCS detection at the document level and employed it to highlight spans of text that exhibit scepticism (Table 4.10). To understand CCS in more detail, we use our best model $\text{MTEXT}_{\text{multi}}$ for NT detection on the examples from Table 4.10 and show them in Table 5.6. Looking at the first example, we see the model categorises it as Condemn as it is blaming the scientific community, and in the second example it uses multiple neutralization techniques in conjunction i.e. Deny-Injury1 and Condemn to construct the argument. Eyeballing them, we can see that these results are in sync, as in first example there is an emphasis on "scientific community" to get its act together and in second example talks about the "hiatus" and deriding "climate scientists" corresponding to Deny-Injury1 and Condemn respectively.

5.4.1 Technical Details

For the supervised BERT models, we use the following fine-tuning hyper-parameters: batch size=10, epoch =3, learning rate=0.0005, number of epochs =3 and use BERT-base-uncased as the base model. We tune our decision boundary threshold to classify the presence of a label based on the development set resulting in 0.2 for Deny-Responsibility, 0.2 for Deny-Injury1, 0.2 for Deny-Injury2, 0.3 for Deny-Victim, 0.3 for Condemn, 0.3 for Loyalties, 0.2 for Justify, and 0.2 for NONE.

For the semi-supervised MTEXT based models, we use the following hyper-parameters: labelled batch size=2, unlabelled batch size=5, sharpening temperature=0.6, the beta distribution parameter = 0.2,[5] learning rate=0.00005, w_1 =1, $w_2 = 1$, $w_3 = 0.8$ and $\alpha = 0.8$, mixing layers= 7,9,12 and BERT model = BERT-base-uncased. We tune the decision boundary threshold to classify the presence of a label based on development set, resulting in 0.75 for Deny-Responsibility, 0.70 for Deny-Injury1, 0.70 for Deny-Injury2, 0.80 for Deny-Victim, 0.85 for Condemn, 0.80 for Loyalties, 0.70 for Justify, and 0.60 for NONE We perform data augmentation for unlabelled data using German and Russian as pivot languages, following Chen et al. (2020).

For SVM, we use unigrams and bigrams as features with tf-idf weighting and the regularization parameter $C = 10$. For $\text{BERT}_{\text{topic}}$ (and $\text{BERT}^{*}_{\text{topic}}$) we use 35 topics similar to Chapter 3, and tune LDA on NSS.

In terms of computing infrastructure, we use RTX 2080 Ti and GTX 1080 GPUs. In MTEXT based models we use 2 GPUs when trained with RTX 2080ti and 3 GPUs when trained with GTX 1080 whereas for BERT-based supervised models they are trained on a single GPU. As all the models are based on BERT-base-uncased, the number of parameters is around 110M.[6] We performed hyperparameter tuning using manual search, based on F-Score on the validation set.

[5]We use a small value here to ensure the generated data in the model is similar to the labelled data with lightweight noise regularization

[6]Strictly speaking number of parameters in $\text{MTEXT}_{\text{multi}}$ is slightly higher due to auxiliary objective, but it is insignificant overall.

5.5 Conclusion

We draw on literature from social science literature and introduced the notion of "neutralisation" in the context of climate change. We developed a dataset made up of CCS sentences from various sources and collected annotations of neutralisation techniques. Next we experimented with supervised pretrained models like `BERT`, domain adapted `BERT`, topic enhanced `BERT` and semi-supervised BERT-based models like MixText. We showed that in instances of sparsely labelled data for a new domain, we get substantial performance gains with the help of unlabelled data. This can be quite useful as collecting large amounts of annotated data is both expensive and time consuming. Additionally, we also see that adding an extra auxiliary supervision signal to predict the frames via a multi-task objective can gives additional small gains. One direction we did not explore in this chapter is the qualitative analysis of error patterns, specifically focusing on human disagreements in annotations and the topics or concepts where these disagreements occur and if the model errors follow similar pattern to human errors. We leave this as part of future work.

So far in the thesis have discussed the detection of climate sceptic articles, identification of highly misleading spans, and identification of neutralization techniques used in the writing. In the next chapter, we will take a different direction by automatically generating explanations, with the help of retrieval augmented generations and knowledge sources, to debunk misleading claims in the climate change domain.

Chapter 6

Automatic Claim Review for Climate Science via Explanation Generation

This chapter builds on:

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "Automatic claim review for climate science via explanation generation." arXiv preprint arXiv:2107.14740 (2021).

This chapter diverges from the paper, with elements such as the problem statement, datasets being retained. While the core problem idea remains the same, this chapter introduces the use of an instruction-tuned LLM of Flan-T5 model, replacing the PLM T5 model used in the paper. Additionally, it delves into the issue of hallucinations in LLMs and explores a reference-free framework based on LLMs to assess the consistency of generated explanations.

6.1 Introduction

The previous two chapters delved into the the task of detecting articles and text spans exhibiting climate change scepticism (CCS), and forming a more nuanced understanding of CCS narratives through the means of framing and neutralization. The natural next step is explaining why or how the claim is inaccurate i.e. the goal is to fact-check claims to

CLAIM

Earth about to enter 30-YEAR 'Mini Ice Age'

VERDICT

INCORRECT

SOURCE: Harry Pettit, Sean Martin, Express, The Sun, 2 Feb. 2020

DETAILS

Factually Inaccurate: The most recent forecast from NOAA's Space Weather Prediction Center (from December 2019) predicts that the next solar cycle will be similar to the one that is currently ending.
Misleading: Even if an extended "grand solar minimum" were to occur, it would not produce marked global cooling.

KEY TAKE AWAY

Scientists cannot predict whether grand solar minimum, which is a decades-long period of lower solar activity, is coming. But even if one occurred, the consequences for average global temperatures would be minimal. Human-caused greenhouse gas emissions will continue to impact average temperatures much more strongly than solar activity cycles.

Figure 6.1: An example of a claim review from climatefeedback.org reproduced from Figure 1.4

verify their truthfulness and give a justification why they may not be truthful. Scientists and experts have been doing this by manually supplying feedback for such claims, verifying their truthfulness and offering the public scientifically sound information. Efforts to fulfill this mission have led to the publication of expert feedback on websites like climatefeedback.org and skepticalscience.com. Figure 6.1 gives us one such example where the claim from The Sun *Earth is about to enter 30-year 'Mini Ice Age'*, has been labelled as Incorrect, with the "Key Take Away" being that *Scientists cannot predict whether solar grand minimum ... is coming* and *even if one occurred, the consequences for average global temperatures would be minimal*. In this chapter, it is this process of fact verification with a textual explanation/justification that we aim to automate, as a tool to assist climate science experts to more efficiently respond to such claims.

We commence this chapter by reviewing issues surrounding claim verification and revisiting Large Language Models (LLMs), as introduced in Chapter 2. Subsequently, we briefly review the literature for retrieval strategies, covering traditional sparse methods like BM25 (Robertson et al., 1995) as well as neural-based techniques like Sentence Embedding (Reimers and Gurevych, 2019b) and Dense Passage Retrieval (Karpukhin et al., 2020). We then delve into the literature on retrieval-augmented generation, tracing its origins to open-domain question-answering systems that employ retriever and reader modules for span detection tasks, and exploring how language models can be integrated with neural retrievers. Next, we review evaluation metrics applied to text generation tasks, which include n-gram-based and semantic-based methods and finally we discuss the idea of "hallucination" in LLMs in detail.

Following the review of the literature, we present a dataset which consists of a knowledge source and the paired data from 'Climate-Fever' which we detailed previously in Section 2.6. We then introduce the task of generating explanations that justify the predicted veracity label for a climate change claim. Next, we detail the architecture and our approach, which draws on work on explainable fact checking (Atanasova et al., 2020) and retrieval-augmented generation (Lewis et al., 2020b), in using the claim to: (1) find documents from a knowledge source such as Wikipedia using a retriever; and (2) generate a veracity label and an explanation for the claim based on the top-k retrieved documents using LLMs like FLan-T5 (Chung et al., 2022).

Thereafter, we experiment with various system configurations and evaluate the generated explanation models not only for their quality when compared to reference explanations but also check for hallucinations i.e. whether the generated explanations are faithful to the retrieved passages. We present results of our experiments and show that our best system can generate high quality explanations. Finally, we employ G-Eval (Liu et al., 2023b), an LLM based reference-free framework to check the consistency of generated explanations, showing that the best generated explanations from our system may be better than the reference explanations.

6.2 Related Work

Fact-checking and the detection of fake news are vital tasks in discussions pertaining to climate science across traditional and social media platforms. Initial research on misinformation predominantly concentrated on areas such as fake news detection (Vlachos and Riedel, 2014), claim and stance verification (Ferreira and Vlachos, 2016), and propaganda detection (Barrón-Cedeno et al., 2019; Da San Martino et al., 2019). Previously in Section 2.6, we looked into the literature concerning general misinformation, fact-checking, and the diverse terminologies related to them. We also discussed the progression of datasets in NLP, such as LIAR (Wang, 2017) — sourced from PolitiFact and categorized into six levels of veracity — and FEVER (Thorne et al., 2018), a dataset generated from Wikipedia with 'supported', 'refuted' and 'not enough info' labels, and its climate change counterpart (Diggelmann et al., 2020). These datasets were examined in conjunction with the stylistic characteristic of misinformation, particularly focusing on the domain of climate science. This chapter specifically zeroes in on the task of claim verification, which assesses the veracity of specific claims using inputs like the claim and contextual information, with more granular output categories reflecting the degree of truthfulness (Hassan et al., 2015; Thorne et al., 2018) and justifications for the assessed veracity. Research closely aligned with explainable fact-checking is found in Atanasova et al. (2020), which uses DistilBERT (Sanh et al., 2019) in a multitask setting and performs the joint task of summarisation and classification of the veracity of the claim.

We divide this section into 2 parts: (1) retrieval augmented generation (Section 6.2.1) and (2) Evaluation (Section 6.2.2). Our focus is on generating explanations for given claims, a process that involves using an external knowledge source to 'retrieve' relevant context for a claim and then 'generate' an explanation based on this information and is thus form of retrieval augmented generation. The task of explanation generation falls within the broader category of text generation similar to tasks like machine translation or summarization, where the generated text is compared with reference texts. To evaluate the quality of generated text with reference texts a range of metrics including n-gram overlap measures like BLEU or ROUGE, and semantic overlap assessments such as BERT-SCORE and BART-SCORE

are used. We will briefly discuss these metrics and also recently introduced reference-free metrics.

6.2.1 Retrieval Augmented Generation

There has recent work on investigating the ability of pretrained models (PLMs) like BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019), and T5 (Raffel et al., 2019) and scaled up LLMs GPT-3/4 (Brown et al., 2020) and Flan–T5 (Chung et al., 2022) to capture factual information (Petroni et al., 2019a). However, "knowledge" in these PLMs and LLMs is stored in the parameters and not directly accessible, making it hard to interpret, extend, or even query these models (Roberts et al., 2020a). One way of augmenting these models is to combine them with external knowledge sources by retrieving passages that are relevant to a given query (a claim, in our case). The text retrieval module can be either traditional keyword based search methods such as BM25 (Robertson et al., 1995), sentence embedding based retrieval (SER) (Reimers and Gurevych, 2019b) or neural retrievers such as dense passage retriever (DPR) (Karpukhin et al., 2020) or its memory efficient counterpart binary passage retriever (BPR) (Yamada et al., 2021). In embedding-based retrieval such as SER, a bi-Encoder architecture comprising two BERT encoders is commonly used to separately map both queries and documents into high-dimensional vector spaces. This model adopts a Siamese network configuration, where the encoder weights are tied and optimized using triplet loss. This loss function serves to minimize the distance between semantically similar query-document pairs while increasing the distance between dissimilar pairs. DPR, like SER, employs a dual-encoder architecture using two distinct BERT models to encode queries and passages. However, a key difference is in the fine-tuning stage, where DPR uses separate, untied weights for each encoder. The relevance score between a query and a passage is computed as the inner product of their respective BERT encodings. An extension of this approach is BPR which incorporates a hashing layer to convert the BERT encodings into binary codes resulting in a more memory-efficient model, without substantial loss in accuracy.

Retrieval-based methods have been successfully applied to open domain question answering (QA) by combining the retriever with a "reader", to extract the relevant answer from those

passages. One of the datasets widely used for for training and evaluating models on tasks of QA is The Stanford Question Answering Dataset (SQuAD). It consists of question-answer pairs based on Wikipedia articles (posed by crowdworkers) where the answer to every question is a segment of text (a span) from the corresponding reading passage. Chen et al. (2017) introduced a span-based extractive framework trained with gold spans in a SQuAD setting (Rajpurkar et al., 2016). The pipeline consists of two components: a document retriever and a document reader, where the document retrieval component employs a TF-IDF-based method to find relevant Wikipedia articles whereas the reader module uses a multi-layer recurrent neural network to process the retrieved documents and extract the answers. Lee and Hsiang (2019) argued against using separate information retrieval systems to retrieve context passages, and proposed "open retrieval question answering"(ORQA), which jointly learns the reader and retriever using only QA pairs (without explicit supervision over context passages). The retriever is pretrained in an unsupervised setting using an "inverse cloze task" i.e. the model is given a specific word or phrase and must identify the context or passage from which it comes. This task allows the retriever to learn how to pick the most contextually relevant passages, providing the necessary background for the query. Similarly, Guu et al. (2020) built on this idea of jointly training a reader and retriever by introducing "Retrieval-augmented language model pre-training "(REALM) which incorporates a learned retriever into language model pre-training, where the model is trained to optimize "salient span masking", a variant of masked language modelling.

Concurrently, Petroni et al. (2020) demonstrated that providing relevant context documents to BERT can dramatically improve its performance on cloze-style factual probing tasks, without requiring any fine-tuning of the model. Following this, Karpukhin et al. (2020) proposed Dense passage Retrieval (DPR) and showed that dense retrieval can outperform sparse BM25 ranking for open-domain Question Answering by learning embeddings optimized for passage ranking. Unlike models like REALM, which integrate retrieval into language modeling, DPR primarily focuses on efficient passage retrieval using a dense vector space. However,though models like ORQA, REALM, and DPR showed promising results they have been limited in scope to open-domain extractive question answering. Addressing this,

Lewis et al. (2020b) proposed retrieval-augmented generation, which combines a pretrained language model with an external knowledge source accessed via a neural retriever such as DPR or BPR, and jointly fine-tuned in a seq2seq manner and excels in a wide range of knowledge-intensive tasks including open domain question answering, abstractive question answering, jeopardy question generation and fact verification. Building on a similar idea, Izacard and Grave (2020a,b) proposed the simple but highly effective "fusion-in decoder" model, which combines evidence from multiple passages independently in separate encoders, and attending to the combined representations in the decoder to generate the answer. Samarinas et al. (2021) extended the idea of passage retrieval to automatic fact checking, and demonstrated that neural retrieval models can improve evidence recall. In our work, we combine these ideas to jointly perform claim veracity classification and generate explanations to justify the prediction.

6.2.2 Evaluation

There is a need to evaluate the quality of generated outputs, which is a challenging task in the broader context of Natural Language Generation. Traditional *n*-gram reference based metrics have been commonly employed for evaluation such as BLEU (Papineni et al., 2002) which emphasizes on precision i.e the presence of generated text *n*-grams in the reference output, and ROUGE (Lin, 2004), focuses on recall i.e. the presence of reference *n*-grams in the generated text. Despite their utility, these metrics fall short in capturing semantic nuances, as they primarily rely on lexical overlap. To address their shortcomings, embedding-based metrics such as BERT-Score (Zhang et al., 2019) and BART-Score (Yuan et al., 2021) have been adopted. BERT-Score leverages contextualized embeddings from the BERT model (Devlin et al., 2019) , and BART-Score employs embeddings from the BART model (Lewis et al., 2020a), to encode both the generated and reference texts into high-dimensional vectors thereby enabling a more nuanced evaluation that accounts for semantic relationships.

Automated metrics offer a valuable approximation for evaluating generated outputs; however, manual annotations remain the gold standard, despite being tedious and expensive. Consequently, there is growing research emphasis on incorporating LLMs into the evaluation

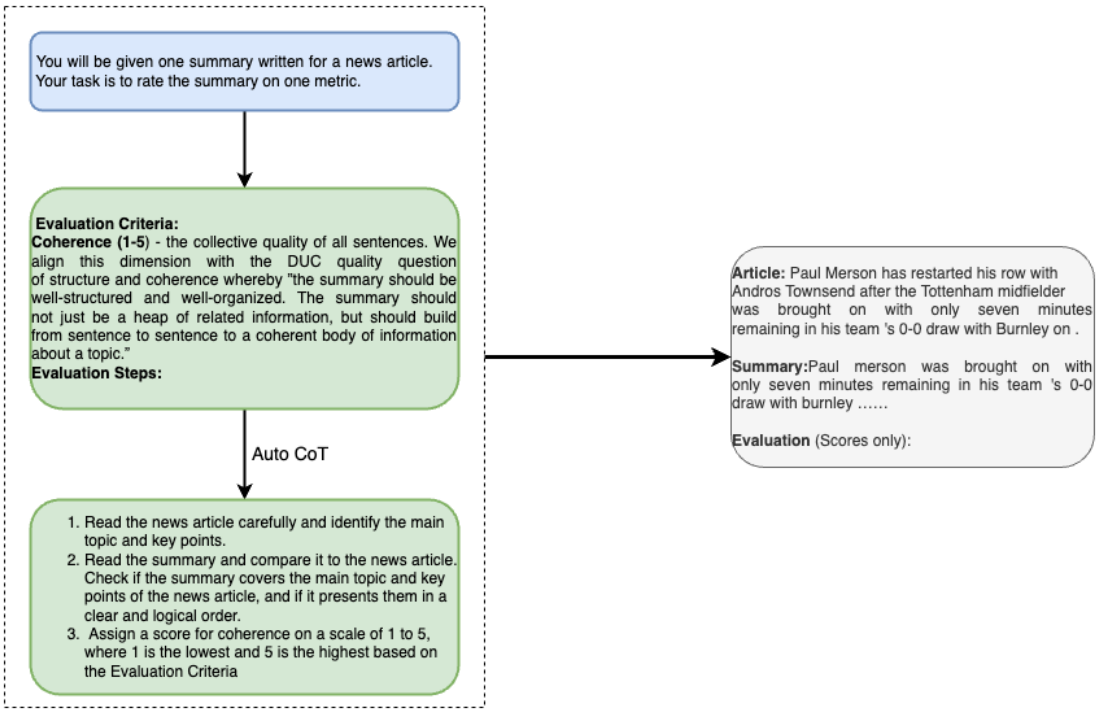


Figure 6.2: G-Eval Auto CoT example reproduced from Liu et al. (2023b)

process to mirror human judgement of generated text in a reference-free evaluation setting. This approach is predicated on the idea that LLMs are capable of scoring outputs based on generative probabilities, underpinning the hypothesis that these models ascribe higher probabilities to texts of higher quality. Fu et al. (2023) introduced GPTScore, leveraging GPT-3 for evaluation through zero-shot instructions and in-context learning. Building on this, Liu et al. (2023b) developed the G-Eval framework, which uses GPT-4 (OpenAI, 2023) to assess the quality of text generation and auto Chain of Thought (CoT) to assess the quality of generated text. This framework incorporates the CoT methodology which was discussed earlier in Section 2.5.2, which prompts the LLM to outline a sequence of intermediate reasoning steps that enhances the evaluation process. The authors demonstrate that LLMs can independently generate these reasoning steps in an “auto CoT” fashion, thereby refining the evaluation results. Figure 6.2 gives an example of this approach, where the LLM is initially fed with the task outline (e.g., summarization in this case) and evaluation criteria. Subsequently, the LLM is instructed to generate auto CoT instructions that specify the

evaluation steps, ultimately applying these steps to score the generated summary in a manner akin to human annotation.

One of the shortcomings in LLMs is that they are prone to "hallucination", which refers to the phenomenon where the generated output contains new information that is not in the provided source content (Filippova, 2020; Maynez et al., 2020). Hallucination can be considered as the converse of "faithfulness", with the latter measures the degree of factual consistency between the generated output and its source. Huang et al. (2021); Ji et al. (2023); Maynez et al. (2020) categorised hallucination into 2 types; intrinsic hallucination, which contradicts the source material; and extrinsic hallucination, where the output introduces content not found in the source. In terms of the underlying causes around hallucination, a primary cause is source-reference divergence i.e. where the training target contains information not present in the source (Ji et al., 2023; Wang, 2020). Parikh et al. (2020) hypothesised that source-reference divergence is not the only reason for hallucinations and there could be other underlying causes such as modelling choices, decoding strategies which increase diversity (Dziri et al., 2021) or parametric knowledge bias in PLMs and LLMs (Petroni et al., 2019b; Roberts et al., 2020b), with Longpre et al. (2021) demonstrating that these models prioritise parametric knowledge over provided input.

In terms of evaluating the degree of hallucination or faithfulness, there has been recent work to study it in the context of abstractive summarization (Durmus et al., 2020; Koto et al., 2022; Wang et al., 2020). Wang et al. (2020) introduced one such framework called QAGS which involves generating questions from a system-generated summary and then answering them based on both the original document and the summary with the factual accuracy is measured by comparing the answers from both sources, using metrics like the F1 score. Koto et al. (2022) highlighted that this QA-based method has limitations, including the need for specific tuning of hyperparameters, high computational demands, and difficulties in adapting it for non-English languages due to specific training data requirements for the QA and question generation models. As an alternative, they proposed using scores from a range of pre-trained models and argued that the best approach to measure faithfulness in abstractive summarization is by comparing the generated summary directly with the source document,

rather than the reference summary, to avoid misclassifying content that is in the source but not in the reference as hallucination. The scoring is given in Equation 6.1

$$\begin{aligned} \text{FA}_{\text{METRIC}} &= \frac{1}{|Y'|} \sum_{t_i \in Y'} A(t_i, X, n) \\ A(t_i, X, n) &= \text{MeanTop-}n\ \text{MET}(t_i, s_j) \quad s_j \in X \end{aligned} \tag{6.1}$$

where t_i and s_j are sentences from the system summary Y' and source document X respectively. MET can be any evaluation method like ROUGE, BERT-SCORE or BART-SCORE and n is a hyperparameter. MeanTop-n matches sentence t_i from the summary with each sentence s_j in the source document X and calculates the average score for the top-n best-matching sentences. Specifically, when $n = 1$ it simplifies to a maximum condition, meaning it only considers the score of the single best-matching sentence. In our context, this means that the generated claim explanations should be assessed for hallucination by comparing them against the retrieved documents rather than claim.

6.3 Datasets

We introduce the datasets in this section. The two key data components of our method are: (1) an external knowledge source ("KS"); and (2) paired claim–explanation data with veracity labels, where the explanation justifies the veracity class.

6.3.1 Knowledge Sources

We commence by introducing the knowledge source, which consist of large-scale databases and document collections. These sources offer an extensive array of information that the LLM leverages to improve its ability to generate responses. Our experiments use Wikipedia ("WIKI") as a primary knowledge source and later in Section 6.7 we employ our best model to examine the impact of altering the knowledge source on generated explanation.

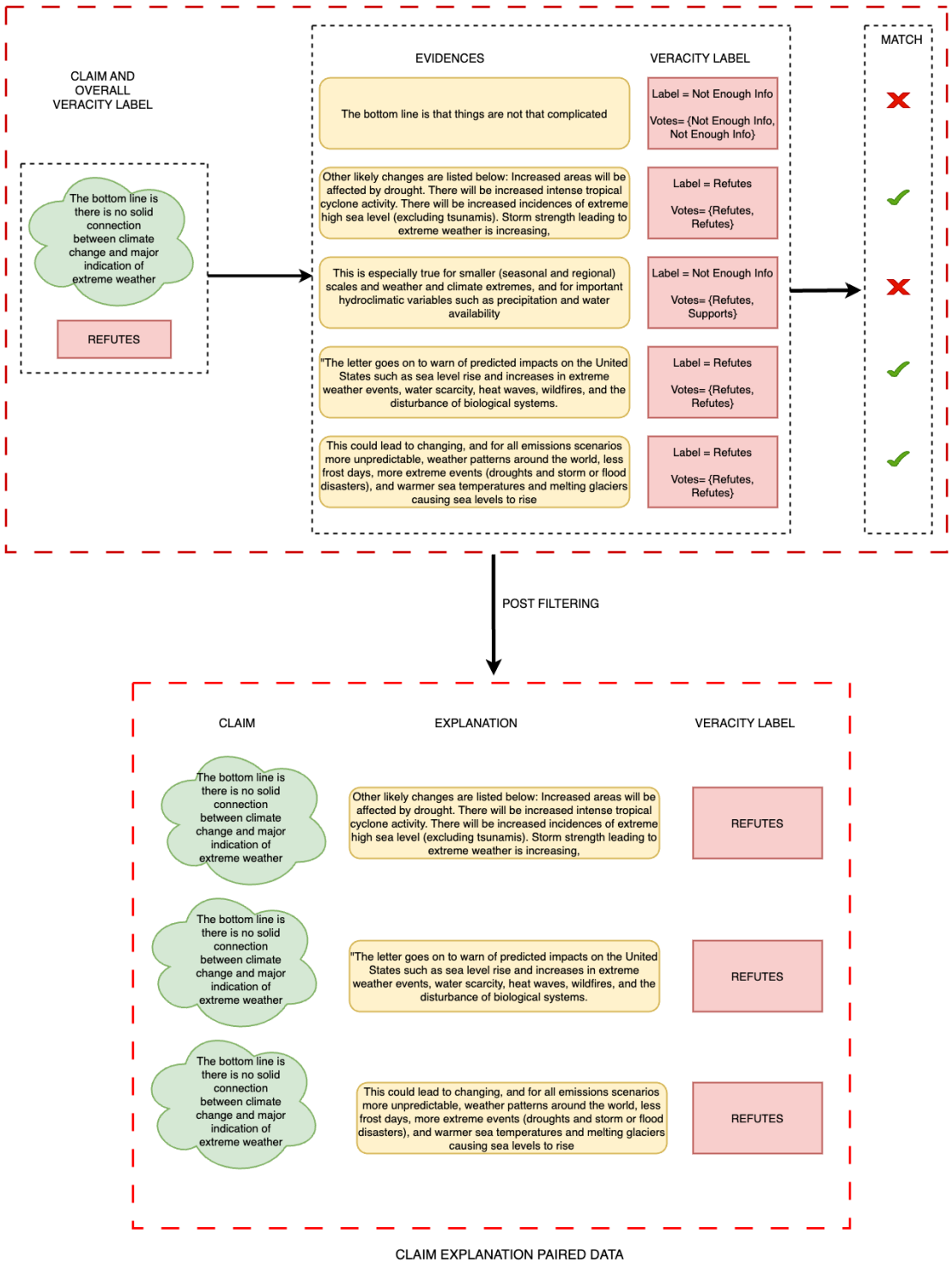


Figure 6.3: Preparation of Claim–Explanation Data

Wikipedia ("WIKI"): We use the processed Wikipedia dump from Dec 2018 as the knowledge source (Chen et al., 2017; Karpukhin et al., 2020).[1] To make the dataset topically aligned to our task of climate change we filter the dataset at article level for the titles that were used to construct evidences in the paired data of CLIMATE-FEVER (see Section 6.3.2). We replicate the preprocessing steps of Chen et al. (2017); Karpukhin et al. (2020) and segment each document into non-overlapping 100-word passages with a minor modification to avoid mid-sentence terminations. To maintain textual integrity, we ensure that passages terminate only at sentence boundaries, thereby avoiding mid-sentence fragmentation. Through this, we generated 74K passages.

6.3.2 Creating Paired Claim—Explanation Data

We previously introduced CLIMATE-FEVER in Section 2.6, and here we simply provide a brief recap as is relevant to our task in this chapter. Diggelmann et al. (2020) released this dataset for climate change claim verification, consisting of 1535 claims, each accompanied by five evidence sentences. Following FEVER (Thorne et al., 2018), it uses Wikipedia as the knowledge source for the evidence sentences, and labels the veracity of each evidence sentence according to 3 classes: supports (`SUP`), refutes (`REF`), and not enough info (`NEI`). Each evidence sentence is annotated by upto 5 annotators, typically 2, and the majority vote determines the micro label for that evidence. In cases of a tie between `SUP` and `REF`, the label defaults to `NEI`. A claim's overall label is based on the majority vote across the evidence labels for its five sentences. However, if there is at least one `SUP` or one `REF` micro label, and the majority is `NEI`, the claim is still assigned either `SUP` or `REF` (whichever has a majority between them). A claim is labelled `NEI` only if all micro labels are `NEI`. The claims serve as the input, and the evidences as the **ground truth explanation or output** for training our explanation generation system. We create a claim-explanation pair for each evidence where its label matches the claim label (micro label matches the overall label) and filter out evidence sentences with differing labels, as shown in Figure 6.3.[2]

[1] Available via the `DPR` repository: https://github.com/facebookresearch/DPR

[2] Strictly speaking, there exists a 'disputed' class with a small proportion of instances, which applies when a claim cannot be definitively categorized into one of the three main classes, often due to a tie in majority votes

Inspired by Lewis et al. (2020b); Thorne and Vlachos (2020) (based on FEVER), we explore 2 configurations of CLIMATE-FEVER in our experiments: (1) 3-way classification ("FEV3 ") i.e. SUP, REF and NEI; and (2) 2-way classification ("FEV2 "), where we consider only SUP vs. REF. We split the two variants of CLIMATE-FEVER into training, validation, and test sets using stratified partitioning. This resulted in: 963 training, 83 validation, and 332 test instances for FEV3; and 680 training, 50 validation, and 177 test instances for FEV2. As each claim has multiple evidence sentences, this translates into a total of 3196 claim-explanation pairs for FEV3, and 1671 claim-explanation pairs for FEV2. In evaluating the quality of generated explanations for a claim, we consider the multiple ground truth explanation sentences as references (in the inference phase). For consistency, we refer to these generated outputs as explanations rather than evidence from this point forward.

Table 6.1 gives an overall example of the claim explanation paired data connected to the knowledge source where the system will take a claim, retrieve the supporting passage from our WIKI knowledge source, and then use this information to generate the explanation and label.

6.4 Method

In this section we describe the model for explanation generation and veracity prediction. Our model is based on the retrieval augmented generation architecture discussed earlier in the chapter (Section 6.2) and uses the Flan-T5 (Chung et al., 2022) – an instruction tuned LLM (Section 2.5.2) —adapted to process claim and support passages, to predict veracity labels and generate explanation texts.

We divide our architecture broadly into two components: (1) retrieval (Section 6.4.1); and (2) explanation and veracity generation (Section 6.4.2). Given a claim c, the role of the retriever is to search for the most relevant (top-k) support passages z_k from a knowledge source (e.g. WIKI). In case of generation, given a claim with k support passages, each support passage z_k is concatenated with the claim c to generate the veracity label and explanation.

between SUP and REF. However, for the sake of consistency with the FEVER literature, we choose to exclude this class from our analysis.

Claim (input)	Short term cooling over the last few years is largely due to a strong La Niña phase in the Pacific Ocean and a prolonged solar minimum
Retrieved Documents	1. Surface temperatures when compared with the average value. The accepted definition is a warming or cooling of at least 0.5 °C (0.9 °F) averaged over the east-central tropical Pacific Ocean. Typically, this anomaly happens at irregular intervals of 2–7 years and lasts nine months to two years. The average period length is 5 years. When this warming or cooling occurs for only seven to nine months, it is classified as El Niño/La Niña "conditions"; when it occurs for more than that period, it is classified as El Niño/La Niña "episodes". 2. During a period of La Niña, the sea surface temperature equatorial Eastern CentralPacific Ocean will be lower than normal by 3–5 °C (5.4–9 °F). An appearance of La Niña often persists for longer than five months. El Niño and La Niña can be indicators of weather changes across the globe. Atlantic and Pacific hurricanes can have different characteristics due to lower or higher wind shear and cooler or warmer sea surface temperatures. 3. This results in changes among ocean currents, and an increase of the subtropical overturning, which is also related to the El Niño and La Niña phenomenon. Depending on stochastic natural variability fluctuations, during La Niña years around 30 % more heat from the upper ocean layer is transported into the deeper ocean. Model studies indicate that ocean currents transport more heat into deeper layers during La Niña years, following changes in wind circulation. Years with increased ocean heat uptake have been associated with negative phases of the interdecadal Pacific oscillation (IPO)
Explanation and Label (output)	During a period of La Niña, the sea surface temperature across the equatorial Eastern Central Pacific Ocean will be lower than normal by 3 to 5°C (5.4 to 9°F) Label: Refutes

Table 6.1: An example of the overall instance with knowledge source and the claim—explanation paired data

The overview of overall architecture is presented in Figure 6.4. It is important to note that the two boxes representing FLAN-T5 do not indicate separate instances of the FLAN T5 model. Instead, they are used to illustrate that the same model (shared parameters) is employed for

two different sub tasks: explanation generation and veracity prediction, each with its own set of instructions.

6.4.1 Retriever: BM25 and SER

We experiment with two retrievers: (1) `BM25` (Robertson et al., 1995); and (2) embedding based retrieval SBERT (`SER`) (Reimers and Gurevych, 2019b). For `BM25`, the knowledge source is stored in the form of document index. Claim texts are tokenised and entities are linked to produce a sparse bag of words/concepts representation. We use Pylucene[3] with default parameters as the retrieval engine, and DBepdia spotlight[4] for entity recognition and linking.

`SER` utilizes a Bi-Encoder architecture featuring two BERT encoders with tied weights, based on a Siamese network structure, to transform queries and documents into high-dimensional vectors. The training objective of `SER` relies on optimizing a triplet loss function consisting of an anchor (query), a positive example (relevant document), and a negative example (irrelevant document). The objective is to minimize the distance between the anchor and the positive example in the high-dimensional space while maximizing the distance between the anchor and the negative example, thereby training the model to cluster semantically similar texts closer together while pushing dissimilar ones apart. During inference both queries and passages are encoded, and candidates are ranked using cosine similarity. We show this in the retrieval part of the architecture in the retriever box part of Figure 6.4. For our experiments, we use "all-mpnet-base-v2"[5] model which is based on the pretrained "mpnet-base"[6] and fine-tuned on a diverse dataset of 1 billion sentence pairs, sourced from a variety of platforms such as Reddit comments (Henderson et al., 2019), S2ORC Citation pairs (Abstracts) (Lo et al., 2020), WikiHow (Koupaee and Wang, 2018), MS MARCO triplets (Nguyen et al., 2016) and SQUAD2 (Rajpurkar et al., 2016) to name a few. SBERT library was used for its implementation.[7]

[3] https://lucene.apache.org/pylucene/
[4] https://www.dbpedia-spotlight.org/api
[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[6] https://huggingface.co/microsoft/mpnet-base
[7] https://www.sbert.net

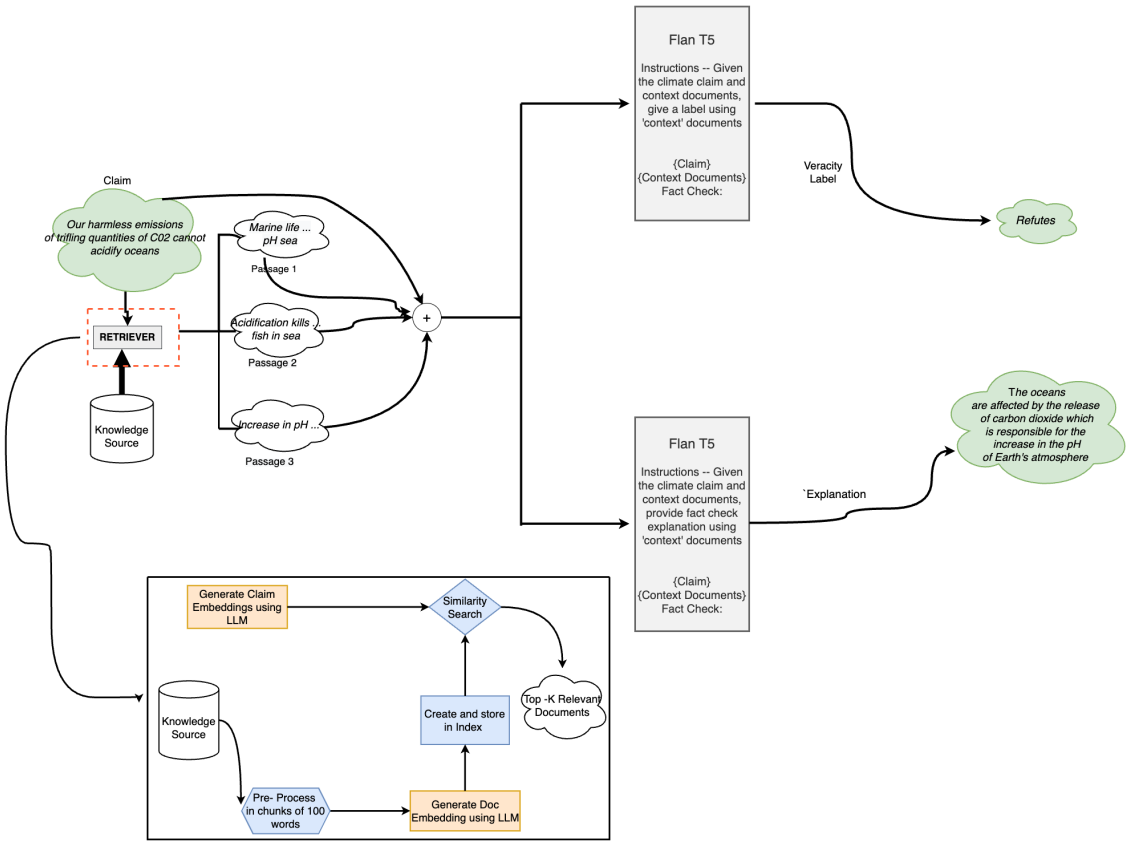


Figure 6.4: Overview of our proposed method for generating an explanation and veracity label for a given claim, based on text passages from a knowledge source.

6.4.2 Explanation Generation

Chung et al. (2022) introduced the Fine-tuned Language Net T5 (FLAN-T5), which refines the capabilities of T5 (Raffel et al., 2019) through the adoption of instruction fine-tuning enabling it to adeptly interpret and execute a diverse array of tasks based on direct language instructions. As discussed earlier in Section 2.5.2, instruction tuning involves preparing instruction formatted instances which includes task descriptions and input-output pairs to fine tune the mode. Here in our specific context, our task description include “Given the climate claim and context documents, provide a fact-check explanation using the context documents" or “Given the climate claim and context documents, provide a veracity label using the context documents" for explanation generation and veracity labelling tasks respectively. The paired claim and explanation data, generated as outlined in Section 6.3.2, serve as the input-output

pairs for this process. We then fine-tune FLAN-T5 using the training partition of this ground truth data. To enhance the model's context understanding, we employ retrieval-augmented generation which involves retrieving relevant passages from a knowledge source using a retriever (Section 6.4.1) based on the input claim and appending this information to the claim, as illustrated in Figure 6.4. For a given claim with k support passages, each support passage z_k is concatenated with the claim c to produce claim–passage contexts $e = [c; z_1; z_2; z_3 ... z_j]$, where $1 \leq j \leq k$. This joint processing of claim—passage (multiple) contexts in the LLM allows the model to summarise the evidence from multiple passages.

6.5 Experiments

In this section, we present the experimental details and evaluation criteria/metrics employed. To reiterate, our experiments use a paired dataset, CLIMATE-FEVER (with evidences items constructed from WIKI) and also use WIKI as the external knowledge source in these experiments.

1. Claim only retrieval (`CL-ONLY`): In this configuration, only the claim is fed into the retrieval system, which then searches for the top-k most relevant passages.

2. Claim and explanation retrieval (`CL-EXP`): In this setting, both the claim and its corresponding paired evidence are input into the retrieval system to search for the top-k relevant passages. The intuition for this approach is that passages crucial for constructing the explanations are more likely to be retrieved to minimize the issue around source—reference divergence. With this approach, it's likely that we could potentially retrieve passages that overlap with the explanation, because these explanations are evidence passages in 'Climate-Fever' which are Wikipedia passages in the first place. It should be noted that this method is employed only during the training phase; for inference, only the claim is used for retrieval (as there is no explanation at test time - that's the output we want to generate)

3. Claim explanation paraphrase retrieval (CL-PP-EXP): Extending the CL-EXP configuration, an additional step is incorporated. Here, the 'paired explanation' is paraphrased using a LLM to maintain semantic coherence, that explanation generation system doesn't learn to simply copy the retrieved passage text at test time after fine-tuned (preventing span copying from the original contextual documents). Similar to CL-EXP, this approach is exclusively applied during the training phase, whereas only claims are used for retrieval during the inference phase.

To assess label veracity prediction, we use classification accuracy (ACC). We evaluate the performance of the generated systems along two axes: (1) the quality of the generated explanations with a given reference explanation; and (2) the faithfulness of these explanations with respect to the retrieved passages. The first axis evaluates the extent to which the generated explanations agree with the reference, and the second measures for hallucinations, ensuring the generated content is grounded by the source (retrieved) passages.

For faithfulness, we borrow the approach of Koto et al. (2022), which we discussed earlier in Section 6.2.2. They assessed the faithfulness in abstractive summarization of a generated summary by comparing it to source documents, where a summary is matched with each source sentence and the average score for the top-*n* best-matching sentences is returned. The intuition behind measuring across the top-*n* is that information in a summary sentence might potentially be drawn from different sentences in the source article. Analogously, we quantify faithfulness by contrasting the generated explanations with the top *k* retrieved passages using ROUGE-1, ROUGE-L, and B-SCORE.[8]

To provide a comprehensive assessment of faithfulness, we introduce two scoring mechanisms: Max score (Max) and Average Top-*k* score (Avg), for the evaluation metrics of B-SCORE, ROUGE-1 and ROUGE-L. The max score is computed based on the highest-scoring passage with the evaluation metrics (B-SCORE, ROUGE-1 and ROUGE-L), while the Average Top-*k* score is calculated as the mean score across the top-*k* retrieved passages. The rationale for employing both metrics is to measure the extent to which multiple passages

[8]Faithfulness and consistency mean the same in this context and are used interchangeably, whereas hallucination is the converse of them

System	Retriever	B-SCORE		ROUGE-1		ROUGE-L		ACC	
		FEV2	FEV3	FEV2	FEV3	FEV2	FEV3	FEV2	FEV3
CL-ONLY	SER	0.85	0.84	0.20	0.21	0.18	0.18	0.78	0.59
	BM25	0.83	0.83	0.20	0.19	0.18	0.18	0.76	0.57
CL-EXP	SER	0.89	0.88	0.30	0.32	0.27	0.26	0.80	0.61
	BM25	0.88	0.87	0.29	0.30	0.24	0.24	0.78	0.58
CL-PP-EXP	SER	0.86	0.86	0.26	0.25	0.20	0.20	0.80	0.61
	BM25	0.86	0.84	0.25	0.24	0.20	0.19	0.79	0.57

Table 6.2: Performance of the models for explanation generation (B-SCORE, ROUGE-1 and ROUGE-L); and veracity prediction (ACC) over WIKI.

contribute to the construction of the explanation. Specifically, a high variance between the Max and the Avg score suggests that the generated explanation is predominantly influenced by a single passage. Conversely, a lower variance indicates that the explanation uses information from multiple retrieved passages.

6.6 Results

In this section we present the empirical findings of our experiments. Our initial analysis compares the efficacy of different retrieval algorithms i.e. BM25 and SER, in the context of veracity prediction and explanation quality. Looking at Table 6.2, we see that SER generally outperforms BM25 for both FEV3 and FEV2. As such, we base the remainder of our experiments exclusively on SER.

We then evaluate the quality of the explanations generated by these models against a reference explanation. We evaluate the performance of explanation generation using ROUGE-1, ROUGE-L (Lin, 2004) and BERT-score (B-SCORE) (Zhang et al., 2019)). As discussed previously in Section 6.2.2, ROUGE-1 and ROUGE-L evaluate the overlap between the generated text and the reference text, and B-SCORE tries to capture the semantic richness of the text, offering a more nuanced evaluation compared to traditional *n*-gram overlap metrics such as ROUGE. Table 6.2 shows that the CL-EXP model consistently performs

the best across the metrics of B-SCORE, ROUGE-1, and ROUGE-L for both FEV3 and FEV2. We hypothesize that the marginally lower performance of the CL-PP-EXP model stems from the inclusion of a paraphrase phase during training, which, may not align as closely with the references as those produced by CL-EXP. Further evaluation of model faithfulness is presented in Table 6.3 and Table 6.4, which, compares performance of different configurations across the Max and Avg scores in the FEV3 and FEV2 setting respectively. For both settings, we see a similar trend. The CL-ONLY configuration scores lowest on both dimensions i.e. performance with respect to a reference explanation and for faithfulness, indicating a propensity towards hallucination. This can attributed to the idea of source divergence bias which we had mentioned previously in Section 6.2.2, wherein the retrieved passages in the training phase does not align with the content of reference evidence claims, prompting the model to learn to hallucinate and generate non-factual content.

Conversely, CL-EXP demonstrates improved scores over CL-ONLY. but with a notable difference between Max and Avg scores. This suggests low degree of hallucination but a narrowed attention to singular passages, thereby reducing the task to copying from one passage. This observation is significant considering that the reference explanations in the CLIMATE-FEVER dataset are also grounded in WIKI, the same source of knowledge for our model. That is, because in the training data we have instances where the retrieved passage overlaps with the explanation, the model is encouraged to do passage copying, leading to this behaviour. Lastly, we observe in CL-PP-EXP an increase in Avg and a reduction in difference between Max and Avg compared to CL-EXP, indicating a more balanced approach. This points to a reduction in hallucination and suggests that the model generates explanations by integrating information from multiple passages.

To check the effect of the number of retrieved documents for both BM25 and SER, we present ACC at different retrieval depths k (between 1 and 20) in Figure 6.5. We see that for both FEV2 and FEV3, in the case of SER we achieve the best performance with 5 documents, before dropping slightly and flattening out. In the case of BM25, it takes more retrieved documents (10) to reach the best performance, before either flattening out or dropping back

System	Dimension	B-SCORE	ROUGE-1	ROUGE-L
CL-ONLY	Max	0.87	0.39	0.25
	Avg	0.83	0.19	0.15
CL-EXP	Max	0.95	0.81	0.78
	Avg	0.84	0.25	0.19
CL-PP-EXP	Max	0.90	0.51	0.46
	Avg	0.86	0.35	0.24

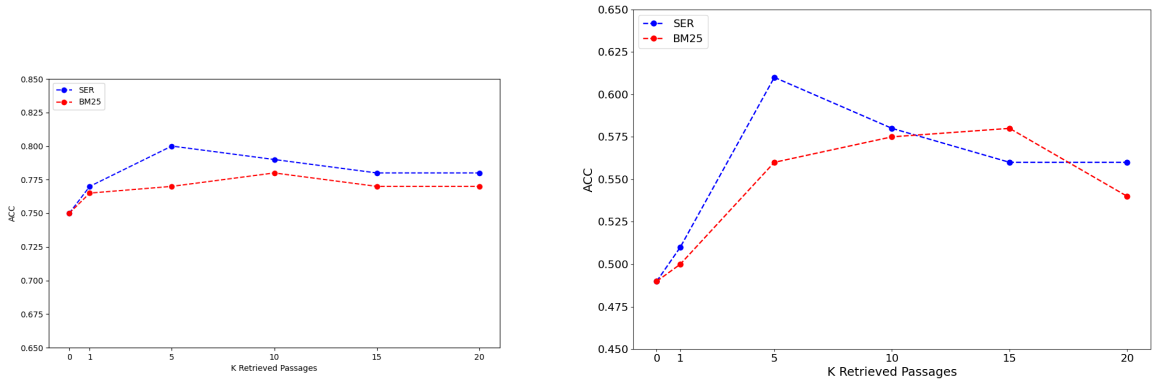
Table 6.3: Performance of the models in terms of faithfulnesses i.e. generation against retrieved documents for FEV3.

System	Dimension	B-SCORE	ROUGE-1	ROUGE-L
CL-ONLY	Max	0.87	0.38	0.24
	Avg	0.84	0.21	0.15
CL-EXP	Max	0.96	0.82	0.78
	Avg	0.86	0.27	0.20
CL-PP-EXP	Max	0.91	0.53	0.47
	Avg	0.86	0.38	0.25

Table 6.4: Performance of the models in terms of faithfulnesses i.e. generation against retrieved documents for FEV2.

in performance, suggesting that the retrieval quality of SER is higher than BM25 for small values of k.

Noting that the generation evaluation metrics (B-SCORE$_{\text{rs}}$, ROUGE-1, and ROUGE-L) may not tell the whole story, we present example generations in Table 6.5 for the three different model configurations. In the case of CL-ONLY, the generation is a hallucinated general statement unfaithful to the the retrieved context whereas in the case of CL-EXP we can see that the generation is a copy of the reference, reinforcing the fact that the system maps into a span detection task in this setting. In CL-PP-EXP, the generated output is longer and draws elements from different retrieved passages, thus resulting in a more coherent and detailed explanation.



(a) FEV2 (b) FEV3

Figure 6.5: ACC performance over different numbers of retrieved documents for SER and BM25, with WIKI.

Text
Claim: Global Warming is not increasing the risk of heatwaves **CL-ONLY**: Global warming is he ongoing increase in global average temperature and its effects on Earth's climate system. **CL-EXP**: Warming boosts the probability of extreme weather events, like heat waves, far more than it boosts more moderate events **CL-PP-EXP**: In the last 30–40 years, heat waves with high humidity have become more frequent and severe, with extreme hot nights having doubled in frequency. These heat waves are caused by the increase in global temperatures, which is causing the Earth's climate to become warmer and more acidic. This is resulting in more frequent heat waves, which can lead to more extreme weather events, such as heat waves. **Reference**: Warming boosts the probability of extreme weather events, like heat waves, far more than it boosts more moderate events

Table 6.5: Example generated explanations for different models with WIKI as knowledge base

6.6.1 LLM - Reference free evaluation

In this section, we use G-Eval framework, a tool devised for reference-free evaluation. We found that CL-PP-EXP produces less hallucination and detailed explanation Table 6.5. But this behaviour (of producing detailed explanation) is penalised by our current reference-based metrics. To this end, we use the G-Eval framework, (Section 6.2) to perform reference-free evaluation to better understand the quality of the generated explanations. It leverages a LLM facilitated by chain-of-thought prompting, enabling a nuanced evaluation of output

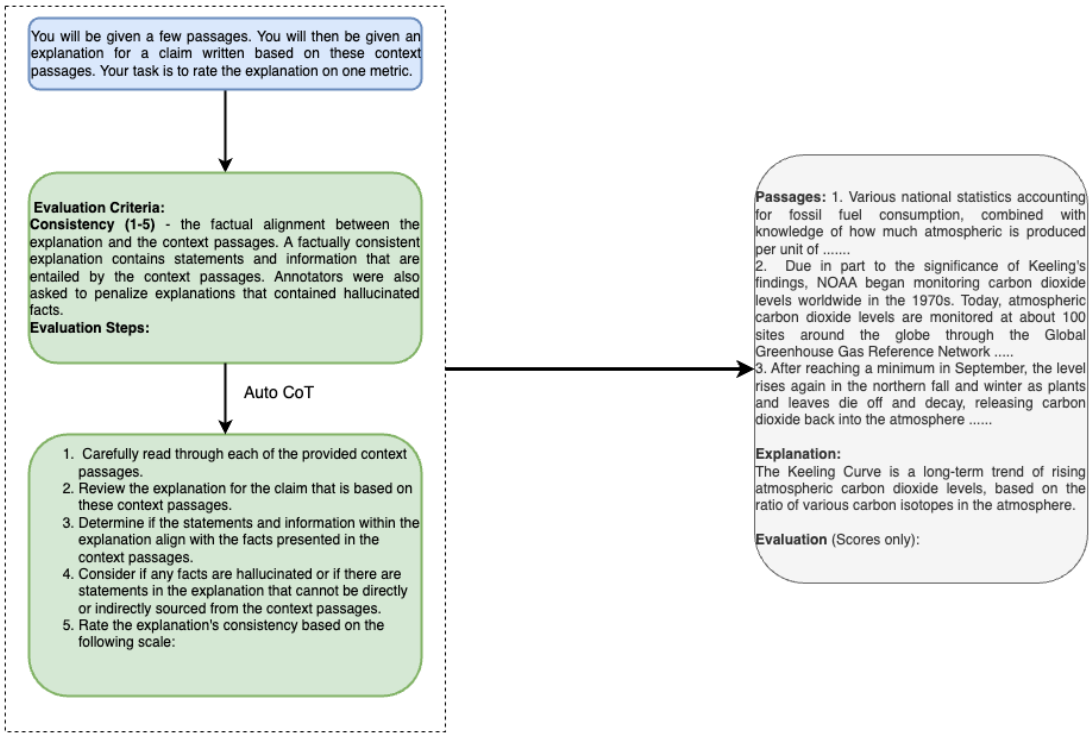


Figure 6.6: GPT-4 Evaluation for Consistency using G-Eval

quality that has shown to correlate strongly with human judgements. We adapt it to our task to check for "consistency" between retrieved passages and reference and generated explanations (by `CL-EXP` and `CL-PP-EXP`). The methodology involves presenting the LLM with a prompt outlining the task, accompanied by evaluation criteria, prompting it to auto-generate a sequence of chain-of-thought steps, as shown in Figure 6.6. In terms of implementation, we use GPT-4 as our LLM of choice and sample 10 iterations with $n = 10$ and a $temperature = 1$, and then calculate the mean score across these samples (analogous to having 10 different annotators).

We present the results of our experiment in Table 6.6. According to the G-Eval framework, `CL-PP-EXP` outperforms other models, demonstrating the fact that `CL-PP-EXP` indeed generates explanations which are grounded better in the retrieved documents. Comparatively, the scores for the reference explanations and those generated by `CL-EXP` are close, often because `CL-EXP` tends to produce explanations with a language similar to the ground truth reference.

Model	Mean Consistency Score
CL-EXP	4.08
CL-PP-EXP	4.58
Reference	4.14

Table 6.6: Consistency evaluation using G-Eval

Text
Claim: CO2 does not cause climate change, it RESPONDS to it ... **PUBS** The rise in global temperature is caused by the increase in atmospheric carbon dioxide (CO2), which is a greenhouse gas that traps longwave radiation and increases the Earth's temperature. This increase in temperature is accompanied by a decrease in the amount of precipitation, which can lead to a rise in sea levels. This can have a significant impact on the environment. Label:REF **CCS** The Earth has been warming for the past 4,000 years, but it has never stopped the Earth from entering an ice age, even when it was 4,000 ppm. This is due to the fact that the Earth is absorbing more and more carbon dioxide, which is a result of the Earth's warming. This means that the earth is able to absorb more and more CO2. Label:SUP **WIKI**: The scientific consensus on climate change is that the Earth is warming at an accelerated rate due to the release of carbon dioxide into the atmosphere. This is a result of the burning of fossil fuels, which releases carbon dioxide directly into the Earth's atmosphere, causing it to warm. This warming is causing the Earth to become warmer, which is resulting in the melting of ice sheets and glaciers. Label:REF
Claim: Currently, sea-level rise does not seem to depend on ocean temperature, and certainly not on CO2 **PUBS**:The rate of sea level rise is expected to continue to increase with continued global warming, and even if temperatures are stabilized through the phasing out of greenhouse gas emissions, sea level is still expected to rise for centuries. This will affect coastal areas worldwide, and robust projections are needed to assess mitigation options and guide adaptation measures. Label:REF **CCS**: The IPCC has consistently refuted the claims of the scientific community that sea levels are rising due to rising temperatures and that the effects of climate change are not yet known and has also argued that the rise in global temperatures is not caused by the increase in greenhouse gases, but rather by the effects on the oceans. This claim is also dependent on the residence time of CO2 in the atmosphere, which probably is much less than alarmists believe. Label:SUP **WIKI**: The Intergovernmental Panel on Climate Change (IPCC) has estimated that we are already committed to a sea-level rise of approximately for each degree of temperature rise within the next 2,000 years. Warming beyond the 2°C target would potentially lead to rates of sea- level rise dominated by ice loss from Antarctica. Continued carbon dioxide emissions from fossil fuel sources could cause additional tens of metres of sea level rise, over the next millennia. Label:REF

Table 6.7: Example generated explanations for the CL-PP-EXP model with PUBS, CCS and WIKI as knowledge bases

6.7 Alternative Knowledge Source?

In the preceding sections, we employed WIKI as the primary knowledge source for our experiments, examining various system configurations to determine optimal performance. However, we did not to assess the impact of alternative knowledge sources may have on explanation generation. Here we will first introduce 2 other knowledge sources on opposite ends of spectrum and then test it on `CL-PP-EXP` system settings.

- **Peer-reviewed PubMed abstracts and IPCC reports ("PUBS"):** A combination of climate change-related abstracts from PubMed,[9] and reports from the Intergovernmental Panel on Climate Change (IPCC).[10] PubMed is a database of peer-reviewed publications primarily in the biomedical domain, but also including other high-profile scientific journals. We sample publications relating to climate science, and extract the title and abstract of each publication. IPCC reports are written by a mix of scientists, experts, and policy makers. They are based off peer-reviewed publications, and are intended to provide a comprehensive summary of a given topic relating to topics such as the physical science of climate change, climate change impacts, or the mitigation of climate change. We apply similar preprocessing on this data as WIKI and segment it into non-overlapping 100-word passages, resulting in 123K passages.

- **Climate Change Scepticism articles ("CCS "):** This is the dataset we complied in Section 4.3 for the task of `CCS` detection from 15 different organizations with known climate change scepticism. The idea behind using this as a knowledge source here is to to test for the impact of an adversarial data source on explanation generation for claims.

Here, we substitute WIKI with the knowledge bases of PUBS and CCS in the `CL-PP-EXP` system without any further fine-tuning. We present examples of generated explanations for different knowledge bases in Table 6.7. Observations from the examples indicate that explanations derived from PUBS have a more scientific undertone, often referencing contexts from

[9] https://pubmed.ncbi.nlm.nih.gov/
[10] https://www.ipcc.ch/

IPCC reports, whereas those from CCS display a spectrum of perspectives, ranging from scepticism to outright misinformation, as evidenced by incorrect citations of IPCC findings in the second example. These differences underscore the nature of the sources within the respective knowledge bases. It is crucial to emphasize that the model exhibits a high degree of faithfulness, employing the retrieved contexts accurately in the generation process but the veracity of the generated explanations turns out to be altogether a different proposition as it is now a function of a knowledge source employed. These results show is that system can generate misleading content, and so the critical thing is to make sure that the knowledge sources are "reputable" (which is not an NLP problem, but rather a meta/world knowledge issue).

6.8 Conclusion

In this chapter, we examined the task of claim verification, focusing on both veracity prediction and explanation generation. We commenced the chapter with a review of retrieval-augmented generation, tracing its roots in open-domain question answering and providing an overview of the literature around the evaluation metrics applied to text generation tasks, and hallucination — examining the underlying causes. We then turned our attention to relevant datasets, such as Climate FEVER and external knowledge sources like Wikipedia setting the stage for our experiments. Subsequently, we detailed our Flan-T5 based retrieval-augmented generation system and experimented with different system settings to highlight the impact of different retrievers, the utility of reference explanations in the retrieval process, and the efficacy of paraphrasing techniques in refining model performance. One minor shortcoming we would like to highlight in the Climate Fever dataset is the use of different units; for instance, claims might use British thermal units while evidence is presented in metric units, leading to discrepancies. Furthermore, the evolving nature of evidence over time means that what may be true today could be invalid in the past or future, a factor not typically highlighted in claims. We did not look into these issues in this chapter but they are important problems that future studies should consider.

We also demonstrated how LLMs can serve as instrumental tools in assessing the quality of explanations generated. While G-Eval looks promising, we do recognize the need for human evaluation, potentially through annotation that involve ranking or direct scoring of explanations. Another direction of human evaluation is to follow the "Fool Me Twice" approach (Eisenschlos et al., 2021), where users assess the veracity of evidence as explanations are presented incrementally, helping to determine which explanations are most effective. Finally we also investigated the changes in generated explanations with different knowledge sources.

Our narrative throughout the chapters have evolved progressively — beginning with the detection of climate change skepticism, understanding its complexities through framing and neutralization, and culminating in this chapter with claim explanation. In the next chapter we summarise our contributions and present avenues for future work.

Chapter 7

Conclusion and Future Work

In this thesis, we have addressed the complex issue of CCS, primarily a blend of misinformation, propaganda, hoaxes, and sensationalism that collectively undermine climate action. This work contributes to the understanding and countering of CCS through the lens of four NLP tasks.

Our first task involved using topic models to extract the underlying themes in documents. We proposed an automated method to optimize and evaluate topic models at the model-document level based on topic-document allocations, thereby helping us enhance the quality of topic outputs that better capture the content of document collection.

Next, we developed a method for detecting CCS articles. By compiling a dataset and applying pre-trained models, we enhanced their ability to identify stylistic and linguistic elements characteristic of CCS, enabling the models to effectively differentiate between CCS and non-CCS articles. We then explored understanding nuances of CCS articles using framing and neutralization techniques. This approach revealed the classes of arguments used in CCS texts. Due to limited annotated data, we applied these NLP techniques in a semi-supervised setting, utilizing unlabelled data to improve classification performance.

Finally, we look into explanation generation, focusing on eliciting the reasons behind a claim's inaccuracy. This involved using LLMs in conjunction with external knowledge bases, such as Wikipedia, to verify facts and ground the generated explanations based on knowledge bases. We also investigated the issue of hallucinations in this context.

Next, we summarise the findings of the thesis, then presents avenues for future work.

7.1 Summary of findings

In **Chapter 3** we introduced an innovative automated method to optimize and evaluate topic models, focusing on the model-document level through examining topic-document allocations. Traditionally, topic models have been optimized using metrics such as perplexity or topic coherence, primarily to rank or filter topics for end-user applications. However, these metrics often provide limited insight into how accurately the topics represent the documents in a collection. To address this gap, we explored an alternative approach for evaluating topic models, centering on topic allocations within documents i.e. via topic intrusion. Our proposed method employs Convolutional Neural Networks with an information retrieval feature vector, to mirror the topic intrusion task. This improved the state-of-the-art at model document-level evaluations and demonstrated the effectiveness of this method in ranking and filtering topics, thereby enhancing the practical utility of topic models in various applications

In **Chapter 4** we introduced the task of climate change scepticism detection, and developed a dataset made up of CCS articles, and documents from a range of non-CCS sources. We proposed novel models based through domain adaptation of PLMs, and demonstrated their effectiveness in both CCS document classification and span detection. We also found that unidirectional models outperformed bidirectional models. Furthermore, we extended our methodology to analyze short texts, such as tweets or brief news excerpts. While this extension yielded some promising results, the performance on short texts did not quite match that of longer texts.

Chapter 5 borrowed concepts from the social science literature and introduced the notion of "neutralisation" and "framing" in the context of climate change scepticism. We developed a dataset comprising of CCS sentences from various sources and annotated them with neutralization and framing classes. Our experimental approach involved the use of both supervised and semi-supervised models. We employed pre-trained models like BERT, along with its domain-adapted and topic-enhanced variants, and explored semi-supervised methods

like MixText. Our findings reveal that in scenarios with limited labeled data in a new domain, significant performance improvements can be achieved by leveraging unlabeled data. This can be quite valuable as collecting large amounts of annotated data is both expensive and time consuming. Furthermore, we also see that incorporating an additional auxiliary supervision signal can yield modest yet noteworthy performance enhancements.

In **Chapter 6** we introduced the task of generating explanations to substantiate the veracity labels assigned to claims about climate change. The chapter began with a review of retrieval-augmented generation, tracing its origins in open-domain question answering, text evaluation and the phenomenon of hallucination. We then discussed relevant datasets, such as Climate FEVER, and external knowledge sources like Wikipedia to lay the groundwork for our experimental investigations. LLMs are adept at encoding real-world knowledge within their parameters and perform well over a range of test understanding and generating tasks. However, they are also prone to producing hallucinations or nonsensical content. To address this, we employed a retrieval-augmented generation approach, wherein the LLM is linked to an external knowledge source through a retriever. This retriever searches the knowledge source for relevant "facts" related to a given claim, and this information is then fed into the LLM, which generates an explanation and veracity label for the claim, grounded in the knowledge source. Our experiments with different retrievers demonstrated that vector-based retrievers outperform keyword-based ones, and also that increasing the number of retrieved documents enhances performance up to a point, peaking at 5 documents. Subsequently, we proposed methodologies to mitigate hallucinations employing strategies such as retrieval using generated explanations in the training phase and paraphrasing. Finally, we demonstrated the utility of LLM-based evaluation methods like G-EVAL to measure the extent of hallucination in the generated explanation and found positive results: the generated explanations are faithful to human written explanations.

7.2 Future Work

Topic Model evaluation at the collection level

In Chapter 3, we proposed an alternative approach to evaluate topic models, focusing on topic allocations in documents through the concept of topic intrusion. This allowed us to move the focus from topic coherence to optimizing topic models based on their topic allocation in each document. However, topic models are also connected to a given collection and evaluation and optimization of topics based on the full collection was not explored.

A good topic model at the collection level should demonstrate a good coverage of the concepts the collection and also distinguish between different collections. For instance, if we apply a topic model to song lyrics from Taylor Swift and The Beatles, the derived topics should clearly differentiate between the two collections, reflecting their distinct characteristics. Similar to the intrusion task discussed in Chapter 3, this could also be conceptualised as an intruder task, where an annotator is presented with a set of topics (e.g. 5 topics) and is tasked to select one topic that does not belong to the collection. However, this task would require individuals (experts) with extensive knowledge of the specific collections, such as the full works of Taylor Swift or The Beatles in this example. Alternatively, we can also train the annotators by giving them time to read through a representative sample of documents in the collection so that they can build an intuition of what should be captured in the topics.

Another approach to evaluating topic models at the collection level could be through the lens of Information Retrieval by framing it as a document retrieval task. Typically, topics in a model are represented as a list of top- n words, but they can also be expressed in other forms, such as phrase labels (Bhatia et al., 2016). Drawing inspiration from the work of Aletras et al. (2017), who evaluated the effectiveness of different modalities (list of n terms, textual labels, or image labels) for a topic in a document retrieval task, we can extend this idea. A well-optimized topic model should be able to retrieve relevant documents effectively. If say a topic, expressed as a list of n terms used for retrieving documents, is too general or coarse-grained, it will retrieve a broad range of documents, many of which may be irrelevant. Conversely, if it is too fine-grained, it may only find a few documents. Therefore, employing a metric that measures the number of relevant documents returned could serve as a means to optimize topics and thus offer a novel method for evaluating topic models at the collection level.

CCS detection for short text

In Chapter 4, we tried to adapt our methodology for detecting CCS in articles by applying it (zero-shot) to short texts, such as tweets. While this approach yielded some promising results, the performance on short texts was notably lower compared to longer texts. This discrepancy highlights a potential area for further research, specifically focusing on optimizing models for the effective detection of CCS in short texts.

Given the limited availability of data for short texts, one viable strategy could involve the generation of synthetic data using LLMs like GPT-4 through zero or a few-shot prompting techniques. Recent studies, such as those by Veselovsky et al. (2023), have demonstrated success in generating faithful synthetic data. Building on this, a PLM could be jointly fine-tuned using both long and short texts. This approach could potentially bridge the gap in CCS detection performance between different text lengths, offering a more robust and versatile model capable of accurately identifying CCS narratives across various text formats.

Neutralization for other domains

In Chapter 5, we introduced the task of neutralization, defined as the justification or vindication for deviant behavior, and proposed a method to categorize CCS arguments into various neutralization technique classes. Our literature review in Chapter 2 highlighted how neutralization has been applied in other domains such as greenwashing, corporate greening, and fast fashion (Joy et al., 2012). Additionally, the tobacco industry's use of neutralization, involved campaigns to undermine scientific standards and spread skepticism (Fooks et al., 2013; Oreskes and Conway, 2010).

Building on our work in the climate domain, an intriguing extension could involve applying these neutralization techniques to other domains. That is, it would be interesting to test whether the model developed for CCS neutralization techniques could be adapted to other contexts, such as a health concern like a pandemic. For instance, the neutralization technique of "appeal to higher loyalties (AHL)" in CCS which prioritizes economic progress over climate action, could be transposed to pandemic scenarios where maintaining economic

stability is deemed more crucial than controlling the spread of the virus. A promising approach could involve fine-tuning an LLM for one setting, such as CCS, and then applying few-shot learning with chain-of-thought (Wei et al., 2022) or tree of thought (Yao et al., 2023) prompting to another context, like a pandemic. This method could potentially enable the transfer of learned neutralization strategies from one domain to another, showcasing the versatility and power of LLMs in understanding and adapting complex argumentative techniques across various fields.

Improving Retrieval Augmented Generation

In Chapter 6, we explored the use of Retrieval Augmented Generation (RAG) with LLMs for verifying and explaining misleading claims to ensure the LLM's explanation is grounded in external knowledge sources. However, this approach has its limitations. The quality of the generated output is heavily dependent on the retrieved results, and thus, on the quality of the retriever itself. Our method uses vector-based embeddings in a bi-encoder setting where two separate encoders independently process a query and a document to retrieve the top k paragraphs and chunks, which are then fed to the LLM for explanation generation. This method is efficient, allowing for the pre-computation and rapid retrieval of document embeddings. One potential improvement to the retriever is the addition of a cross-encoder reranker which uses a single encoder to jointly process both the query and the document. This joint processing allows the encoder to consider the interaction between the query and document, which can lead to more accurate and context-aware relevance assessments. As a cross-encoder is computationally expensive, one solution could be to use a fast retriever to get an initial set of candidates before using the cross-encoder as a reranker.

Another challenge is the excess of irrelevant information in the retrieved paragraphs. To address this, we could take the claim and the retrieved paragraphs from the retriever and "compress" each paragraph by retaining only the context relevant to the claim. A separate LLM could be trained specifically for this task. Recent work by Liu et al. (2023a) shows that LLMs tend to miss information in the middle of longer contexts, making this strategy crucial to ensure that only pertinent parts are retained. To further ensure the fidelity of the generated

output to the source, we could make the model cite the paragraphs it used for generating the output. Another direction, as explored in recent work by Yue et al. (2023), involves the automatic evaluation of attribution. Such method could be used to ensure that the generated output is supported by the cited reference, but also verify that the cited reference itself is not a product of hallucination.

Bibliography

Abram, N. J., Henley, B. J., Sen Gupta, A., Lippmann, T. J., Clarke, H., Dowdy, A. J., Sharples, J. J., Nolan, R. H., Zhang, T., Wooster, M. J., et al. (2021). Connections of climate change and variability to large and extreme forest fires in southeast australia. *Communications Earth & Environment*, 2(1):1–17.

Aletras, N., Baldwin, T., Lau, J. H., and Stevenson, M. (2017). Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1):154–167.

Aletras, N. and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the Tenth International Workshop on Computational Semantics (IWCS-10)*, pages 13–22, Potsdam, Germany.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Anderegg, W. R., Prall, J. W., Harold, J., and Schneider, S. H. (2010). Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, 107(27):12107–12109.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Arrhenius, S. (1896). Xxxi. on the influence of carbonic acid in the air upon the temperature of the ground. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(251):237–276.

Arrhenius, S. (2011). On the influence of carbonic acid in the air upon the temperature of the ground. *The Warming Papers: The Scientific Foundation for the Climate Change Forecast*, pages 56–77.

Atanasova, D. and Koteyko, N. (2017). Metaphors in guardian online and mail online opinion-page content on climate change: War, religion, and politics. *Environmental Communication*, 11(4):452–469.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020). Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364.

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Baird, H. S. (2007). The state of the art of document image degradation modelling. *Digital Document Processing*, pages 261–279.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Balcan, M.-F., Blum, A., and Yang, K. (2004). Co-training and expansion: Towards bridging theory and practice. *Advances in neural information processing systems*, 17.

Barrón-Cedeno, A., Da San Martino, G., Jaradat, I., and Nakov, P. (2019). Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Benegal, S. D. and Scruggs, L. A. (2018). Correcting misinformation about climate change: The impact of partisanship in an experimental setting. *Climatic change*, 148(1):61–80.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

Benson, R. (2013). *Shaping immigration news*. Cambridge University Press Cambridge.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Bhatia, S., Lau, J. H., and Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963.

Bhatia, S., Lau, J. H., and Baldwin, T. (2017). An automatic approach for document-level topic model evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 206–215. Association for Computational Linguistics.

Biyani, P., Tsioutsiouliklis, K., and Blackmer, J. (2016). "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Blei, D. and Lafferty, J. (2006a). Correlated topic models. *Advances in neural information processing systems*, 18:147.

Blei, D. and Lafferty, J. (2006b). Correlated topic models. *Advances in Neural Information Processing Systems*, 18.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Blei, D. M. and Lafferty, J. D. (2006c). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blum, A., Lafferty, J., Rwebangira, M. R., and Reddy, R. (2004). Semi-supervised learning using randomized mincuts. In *Proceedings of the twenty-first international conference on Machine learning*, page 13.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Boussalis, C. and Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change*, 36:89–100.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2013). Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop, London*.

Breshears, D. D., Cobb, N. S., Rich, P. M., Price, K. P., Allen, C. D., Balice, R. G., Romme, W. H., Kastens, J. H., Floyd, M. L., Belnap, J., et al. (2005). Regional vegetation die-off in response to global-change-type drought. *Proceedings of the National Academy of Sciences*, 102(42):15144–15148.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bryson, R. A. (1974). A perspective on climatic change: Climate responds rapidly and significantly to small changes of the independent variables. *Science*, 184(4138):753–760.

Buckels, E. E., Trapnell, P. D., and Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and individual Differences*, 67:97–102.

Buntine, W. L. and Mishra, S. (2014). Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 881–890.

Burnard, L. (1995). User guide for the British National Corpus.

Callendar, G. S. (1949). Can carbon dioxide influence climate? *Weather*, 4(10):310–314.

Cao, Z., Li, S., Liu, Y., Li, W., and Ji, H. (2015). A novel neural topic model and its supervised extension. In *Proceedings of AAAI 2015*, pages 2210–2216.

Card, D., Gross, J. H., Boydstun, A., and Smith, N. A. (2016). Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1410–1420.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 21 (NIPS-09)*, pages 288–296, Vancouver, Canada.

Chen, C., Du, L., and Buntine, W. (2011). Sampling table configurations for the hierarchical poisson-dirichlet process. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 296–311. Springer.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Chen, J., Yang, Z., and Yang, D. (2020). MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Cheng, X., Shuai, C., Liu, J., Wang, J., Liu, Y., Li, W., and Shuai, J. (2018). Topic modelling of ecology, environment and poverty nexus: An integrated framework. *Agriculture, ecosystems & environment*, 267:1–14.

Chillrud, G. W. L. and McKeown, K. (2021). Evidence based automatic fact-checking for climate change misinformation.

Chong, D. and Druckman, J. N. (2007). Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Clark, K., Luong, M.-T., Manning, C. D., and Le, Q. (2018). Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., Way, R., Jacobs, P., and Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental research letters*, 8(2):024024.

Croce, D., Castellucci, G., and Basili, R. (2020). GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119.

Da San Martino, G., Yu, S., Barrón-Cedeno, A., Petrov, R., and Nakov, P. (2019). Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5640–5650.

Dahal, B., Kumar, S. A., and Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. (2019). Plug and play language models: a simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Delmas, M. A. and Burbano, V. C. (2011). The drivers of greenwashing. *California management review*, 54(1):64–87.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Diethelm, P. and McKee, M. (2009). Denialism: what is it and how should scientists respond? *The European Journal of Public Health*, 19(1):2–4.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. (2020). Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Dowdy, A. J., Ye, H., Pepler, A., Thatcher, M., Osbrough, S. L., Evans, J. P., Di Virgilio, G., and McCarthy, N. (2019). Future changes in extreme weather and pyroconvection risk factors for australian wildfires. *Scientific reports*, 9(1):1–11.

Doyle, G. and Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 281–288.

Dunlap, R. E. and Brulle, R. J. (2015). *Climate change and society: Sociological perspectives*. Oxford University Press.

Dunlap, R. E. and Jacques, P. J. (2013). Climate change denial books and conservative think tanks: Exploring the connection. *American Behavioral Scientist*, 57(6):699–731.

Durmus, E., He, H., and Diab, M. (2020). Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Dziri, N., Madotto, A., Zaiane, O., and Bose, A. J. (2021). Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Eisenschlos, J., Dhingra, B., Bulian, J., Börschinger, B., and Boyd-Graber, J. (2021). Fool me twice: Entailment from wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ekholm, N. (1901). On the variations of the climate of the geological and historical past and their causes. *Quarterly Journal of the Royal Meteorological Society*, 27(117):1–62.

Elgesem, D., Steskal, L., and Diakopoulos, N. (2019). Structure and content of the discourse on climate change in the blogosphere: The big picture. In *Climate Change Communication and the Internet*, pages 21–40. Routledge.

Elsasser, S. W. and Dunlap, R. E. (2013). Leading voices in the denier choir: Conservative columnists' dismissal of global warming and denigration of climate science. *American Behavioral Scientist*, 57(6):754–776.

Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, pages 390–397.

Farrell, J. (2016). Network structure and influence of the climate change counter-movement. *Nature Climate Change*, 6(4):370.

Farrell, J. (2019). The growth of climate change misinformation in us philanthropy: evidence from natural language processing. *Environmental Research Letters*, 14(3):034013.

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

Filippova, K. (2020). Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Fleming, J. (2013). *The callendar effect: the life and work of Guy Stewart Callendar (1898-1964)*. Springer Science & Business Media.

Fleming, J. R. (2005). *Historical perspectives on climate change*. Oxford University Press.

Fløttum, K. (2014). Linguistic mediation of climate change discourse. *ASp. la revue du GERAS*, (65):7–20.

Flottum, K. (2017). *The role of language in the climate change debate*. Taylor & Francis.

Fløttum, K., Dahl, T., and Rivenes, V. (2016). Young Norwegians and their views on climate change and the future: findings from a climate concerned and oil-rich nation. *Journal of Youth Studies*, 19(8):1128–1143.

Fooks, G., Gilmore, A., Collin, J., Holden, C., and Lee, K. (2013). The limits of corporate social responsibility: techniques of neutralization, stakeholder management and political CSR. *Journal of Business Ethics*, 112(2):283–299.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2023). Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Gardner, A. S., Moholdt, G., Scambos, T., Fahnstock, M., Ligtenberg, S., Van Den Broeke, M., and Nilsson, J. (2018). Increased west antarctic and unchanged east antarctic ice discharge over the last 7 years. *The Cryosphere*, 12(2):521–547.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Goss, M., Swain, D. L., Abatzoglou, J. T., Sarhadi, A., Kolden, C. A., Williams, A. P., and Diffenbaugh, N. S. (2020). Climate change is increasing the likelihood of extreme autumn wildfire conditions across california. *Environmental Research Letters*, 15(9):094016.

Grandvalet, Y. and Bengio, Y. (2005). Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

Gururangan, S., Dang, T., Card, D., and Smith, N. A. (2019). Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5880–5894.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020). Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Hansen, J. (1988). Statement of dr. james hansen, director, nasa goddard institute for space studies. *Congressional Record, June*, 23:1988.

Harrison, S., Kargel, J. S., Huggel, C., Reynolds, J., Shugar, D. H., Betts, R. A., Emmer, A., Glasser, N., Haritashya, U. K., Klimeš, J., et al. (2018). Climate change and the global pattern of moraine-dammed glacial lake outburst floods. *The Cryosphere*, 12(4):1195–1209.

Hassan, N., Li, C., and Tremayne, M. (2015). Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th acm international on conference on information and knowledge management*, pages 1835–1838.

Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrkšić, N., Spithourakis, G., Su, P.-H., Vulić, I., et al. (2019). A repository of conversational datasets. *arXiv preprint arXiv:1904.06472*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Horne, B. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Huang, Y., Feng, X., Feng, X., and Qin, B. (2021). The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Ito, R., Nakae, K., Hata, J., Okano, H., and Ishii, S. (2019). Semi-supervised deep learning of brain tissue segmentation. *Neural Networks*, 116:25–34.

Iyengar, S. (1994). *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., et al. (2022). Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Izacard, G. and Grave, E. (2020a). Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

Izacard, G. and Grave, E. (2020b). Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jiang, S. and Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23.

Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226.

Joachims, T. et al. (1999). Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209.

Jones, M. W., Smith, A., Betts, R., Canadell, J. G., Prentice, I. C., and Le Quéré, C. (2020). Climate change increases the risk of wildfires. *ScienceBrief Review*, 116:117.

Joty, S. and Mohiuddin, M. T. (2018). Modeling speech acts in asynchronous conversations: A neural-crf approach. *Computational Linguistics*, 44(4):859–894.

Jowett, G. S. and O'donnell, V. (2018). *Propaganda & persuasion*. Sage publications.

Joy, A., Sherry Jr, J. F., Venkatesh, A., Wang, J., and Chan, R. (2012). Fast fashion, sustainability, and the ethical appeal of luxury brands. *Fashion Theory*, 16(3):273–295.

Jwa, H., Oh, D., Park, K., Kang, J. M., and Lim, H. (2019). exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, 9(19):4062.

Kaliyar, R. K., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Kaptein, M. and Van Helvoort, M. (2019). A model of neutralization techniques. *Deviant Behavior*, 40(10):1260–1285.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Keeling, C. D. (1960). The concentration and isotopic abundances of carbon dioxide in the atmosphere. *Tellus*, 12(2):200–203.

Keeling, C. D. (1998). Rewards and penalties of monitoring the earth. *Annual Review of Energy and the Environment*, 23(1):25–82.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *EMNLP*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klockars, C. B. J. (1974). *The Professional Fence*. The Free Press.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Koto, F., Baldwin, T., and Lau, J. H. (2022). Ffci: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, 73:1553–1607.

Koupaee, M. and Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Kreienkamp, F., Philip, S. Y., Tradowsky, J. S., Kew, S. F., Lorenz, P., Arrighi, J., Belleflamme, A., Bettmann, T., Caluwaerts, S., Chan, S. C., et al. (2021). Rapid attribution of heavy rainfall events leading to the severe flooding in western europe during july 2021.

Kukla, G. J. and Kočí, A. (1972). End of the last interglacial in the loess record1. *Quaternary Research*, 2(3):374–383.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., and Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, 62(3):790–816.

Kumar, A., Bhattamishra, S., Bhandari, M., and Talukdar, P. (2019). Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619.

Kumar, S., West, R., and Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602.

Kwak, H., An, J., and Ahn, Y.-Y. (2020). A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *12th ACM Conference on Web Science*, pages 305–314.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lau, J. H., Armendariz, C., Lappin, S., Purver, M., and Shu, C. (2020). How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Lau, J. H., Armendariz, C. S., Lappin, S., Purver, M., and Shu, C. (2020). How furiously can colourless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, pages 296–310.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of EACL 2014*, pages 530–539.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Lee, J.-S. and Hsiang, J. (2019). Patent claim generation by fine-tuning openai gpt-2. *arXiv preprint arXiv:1907.02052*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020a). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Lewis, O. T. (2006). Climate change, species–area curves and the extinction crisis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1465):163–171.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020b). Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Li, J. J. and Nenkova, A. (2015). Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., and Lu, X. (2019a). A two-stage model based on bert for short fake news detection. In *Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, August 28–30, 2019, Proceedings, Part II 12*, pages 172–183. Springer.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023a). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023b). Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S. (2020). S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., and Singh, S. (2021). Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Luo, Y., Card, D., and Jurafsky, D. (2020). DeSMOG: Detecting stance in media on global warming. In *Findings of EMNLP 2020*.

Lynch, M. J., Burns, R. G., and Stretesky, P. B. (2010). Global warming and state-corporate crime: the politicalization of global warming under the bush administration. *Crime, Law and Social Change*, 54(3):213–239.

Lynch, M. J. and Stretesky, P. (2013). Green criminology in the united states. In *Issues in green criminology*, pages 270–291. Willan.

MacKay, D. J. (2016). *Sustainable Energy-without the hot air*. Bloomsbury Publishing.

Manabe, S. and Bryan, K. (1969). Climate calculations with a combined ocean-atmosphere model. *J. Atmos. Sci*, 26(4):786–789.

Manabe, S. and Wetherald, R. T. (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity.

Maruna, S. and Copes, H. (2005). What have we learned from five decades of neutralization research? *Crime and justice*, 32:221–320.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of artificial intelligence research*, 30:249–272.

McCright, A. M. and Dunlap, R. E. (2000). Challenging global warming as a social problem: An analysis of the conservative movement's counter-claims. *Social problems*, 47(4):499–522.

McKie, R. (2018). *Rebranding the Climate Change Counter Movement through a Criminological and Political Economic Lens*. PhD thesis, Northumbria University.

Mendelsohn, J., Budak, C., and Jurgens, D. (2021). Modeling framing in immigration discourse on social media. *arXiv preprint arXiv:2104.06443*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. pages 746–751, Atlanta, USA.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 262–272, Edinburgh, UK.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., and Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Minor, W. W. (1981). Techniques of neutralization: A reconceptualization and empirical examination. *Journal of Research in Crime and Delinquency*, 18(2):295–318.

Mohr, J. W. and Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2008). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Moser, S. C. (2010). Communicating climate change: history, challenges, process and future directions. *Wiley Interdisciplinary Reviews: Climate Change*, 1(1):31–53.

Nakamura, K., Levy, S., and Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Nicholls, N. (2007). Climate: Sawyer predicted rate of warming in 1972. *Nature*, 448(7157):992–992.

OpenAI (2023). Gpt-4 technical report.

Oreskes, N. and Conway, E. M. (2010). Defeating the merchants of doubt. *Nature*, 465(7299):686–687.

Oreskes, N. and Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Parikh, A. P., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., and Das, D. (2020). Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Paul, M. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 265–272.

Pauls, A. and Klein, D. (2012). Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968.

Peinelt, N., Nguyen, D., and Liakata, M. (2020). tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055.

Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296.

Pennington, J., Socher, R., and Manning, C. (2014a). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pennington, J., Socher, R., and Manning, C. D. (2014b). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018a). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018b). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., and Riedel, S. (2020). How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019a). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019b). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Pomper, G. M. and Lederman, S. S. (1980). *Elections in America: Control and influence in democratic politics*. Longman Publishing Group.

Przybyla, P. (2020). Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 490–497.

Qazvinian, V., Rosengren, E., Radev, D., and Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1589–1599.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Rabitz, F., Telešienė, A., and Zolubienė, E. (2021). Topic modelling the news media representation of climate change. *Environmental Sociology*, 7(3):214–224.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ranney, M. A. and Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in cognitive science*, 8(1):49–75.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Raza, S. and Ding, C. (2022). Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362.

Reimers, N. and Gurevych, I. (2019a). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reimers, N. and Gurevych, I. (2019b). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Revelle, R. and Suess, H. E. (1957). Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric co2 during the past decades. *Tellus*, 9(1):18–27.

Rhodes, C. J. (2016). The 2015 paris climate change conference: Cop21. *Science progress*, 99(1):97–104.

Roberts, A., Raffel, C., and Shazeer, N. (2020a). How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Roberts, A., Raffel, C., and Shazeer, N. (2020b). How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Robertson, S. and Walker, S. (1994). Some simple effective approximations to the 2-Poisson Model for probabilistic weighted retrieval. In *Proceedings of 17th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 232–241.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015)*, pages 399–408, Shanghai, China.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005). Semi-supervised self-training of object detection models.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.

Salway, A., Touileb, S., and Tvinnereim, E. (2014). Inducing information structures for data-driven text analysis. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 28–32.

Samarinas, C., Hsu, W., and Lee, M. L. (2021). Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sawyer, J. S. (1972). Man-made carbon dioxide and the "greenhouse" effect. *Nature*, 239(5366):23–26.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. (2000). Support vector method for novelty detection. In *Advances in neural information processing systems*, pages 582–588.

Searle, J. (1976). A taxonomy of illocutionary acts (pp. 355-68). *Linguistic Agency University of Trier*.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Severyn, A. and Moschitti, A. (2015). Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 373–382. ACM.

Sharman, A. (2014). Mapping the climate sceptical blogosphere. *Global Environmental Change*, 26:159–170.

Sherwood, S. (2011). Science controversies past and present. *Physics Today*, 64(10):39.

Shigihara, A. M. (2013). It's only stealing a little a lot: Techniques of neutralization for theft among restaurant workers. *Deviant Behavior*, 34(6):494–512.

Shin, J., Lee, Y., and Jung, K. (2019). Effective sentence scoring method using BERT for speech recognition. In *Asian Conference on Machine Learning*, pages 1081–1093.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Sindhwani, V., Niyogi, P., and Belkin, M. (2005). A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML workshop on learning with multiple views*, volume 2005, pages 74–79. Citeseer.

Sindhwani, V. and Rosenberg, D. S. (2008). An rkhs for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, pages 976–983.

Sleeman, J., Halem, M., Finin, T., and Cane, M. (2017). Modeling the evolution of climate change assessment research using dynamic topic models and cross-domain divergence maps. In *2017 AAAI Spring Symposium Series*.

Stammbach, D. and Ash, E. (2020). e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

Stehr, N., Von Storch, H., and Flügel, M. (1995). The 19th century discussion of climate variability and climate change: Analogies for the present debate? *World Resource Review*, 7:589–604.

Subramanian, S., Cohn, T., and Baldwin, T. (2019a). Deep ordinal regression for pledge specificity prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1729–1740.

Subramanian, S., Cohn, T., and Baldwin, T. (2019b). Target based speech act classification in political campaign text. *NAACL HLT 2019*, page 273.

Sykes, G. M. and Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *American sociological review*, 22(6):664–670.

Targ, S., Almeida, D., and Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

Teh, Y., Newman, D., and Welling, M. (2006). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Advances in neural information processing systems*, 19.

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, 74(4):905–916.

Thorne, J. and Vlachos, A. (2020). Avoiding catastrophic forgetting in mitigating model biases in sentence-pair classification with elastic weight consolidation. *arXiv preprint arXiv:2004.14366*.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Triguero, I., García, S., and Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284.

Tvinnereim, E. and Fløttum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature Climate Change*, 5(8):744–747.

Tyndall, J. (1872). *Contributions to molecular physics in the domain of radiant heat*. Longmans, Green and Company.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. (2016). Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., and Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2):1600008.

Van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

van Oldenborgh, G. J., Krikken, F., Lewis, S., Leach, N. J., Lehner, F., Saunders, K. R., van Weele, M., Haustein, K., Li, S., Wallom, D., Sparrow, S., Arrighi, J., Singh, R. P., van Aalst, M. K., Philip, S. Y., Vautard, R., and Otto, F. E. L. (2020). Attribution of the australian bushfire risk to anthropogenic climate change. *Natural Hazards and Earth System Sciences Discussions*, 2020:1–46.

Varini, F. S., Boyd-Graber, J., Ciaramita, M., and Leippold, M. (2020). Climatext: A dataset for climate change topic detection. *arXiv preprint arXiv:2012.00483*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Bengio, Y., and Lopez-Paz, D. (2019). Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*.

Veselovsky, V., Ribeiro, M. H., Arora, A., Josifoski, M., Anderson, A., and West, R. (2023). Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Vijjali, R., Potluri, P., Kumar, S., and Teki, S. (2020). Two stage transformer model for covid-19 fake news detection and fact checking. *arXiv preprint arXiv:2011.13253*.

Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

Vraga, E. K. and Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1):136–144.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, A. and Cho, K. (2019). BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Wang, F. and Zhang, C. (2006). Label propagation through linear neighborhoods. In *Proceedings of the 23rd international conference on Machine learning*, pages 985–992.

Wang, H. (2020). Revisiting challenges in data-to-text generation with fact grounding. *arXiv preprint arXiv:2001.03830*.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Washington, H. (2013). *Climate change denial: Heads in the sand*. Routledge.

Weart, S. R. (2010). The idea of anthropogenic global climate change in the 20th century. *Wiley Interdisciplinary Reviews: Climate Change*, 1(1):67–81.

Webersinke, N., Kraus, M., Bingler, J. A., and Leippold, M. (2021). Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wu, L., Morstatter, F., Carley, K. M., and Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. (2020a). Unsupervised data augmentation for consistency training. In *Proceedings of NeurIPS 2020*.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020b). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Yamada, I., Asai, A., and Hajishirzi, H. (2021). Efficient passage retrieval with hashing for open-domain question answering. *arXiv preprint arXiv:2106.00882*.

Yang, Y., Nan, F., Yang, P., Meng, Q., Xie, Y., Zhang, D., and Muhammad, K. (2019a). GAN-based semi-supervised learning approach for clinical decision support in health-iot platform. *IEEE Access*, 7:8048–8057.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019b). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yang, Z., Hu, Z., Salakhutdinov, R., and Berg-Kirkpatrick, T. (2017). Improved variational autoencoders for text modeling using dilated convolutions. In *International Conference on Machine Learning*, pages 3881–3890.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Yuan, W., Neubig, G., and Liu, P. (2021). Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Yue, X., Wang, B., Zhang, K., Chen, Z., Su, Y., and Sun, H. (2023). Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, R., Gao, D., and Li, W. (2012). Towards scalable speech act recognition in twitter: tackling insufficient training data. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 18–27.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhou, D., Bousquet, O., Lal, T., Weston, J., and Schölkopf, B. (2003). Learning with local and global consistency. *Advances in neural information processing systems*, 16.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919.