Thesis Proposal

Automated Misinformation Detection with Natural Language Processing and Network Science Models

Ian Kloo

April 2024

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kathleen M. Carley, Chair, Carnegie Mellon University
Dr. Brandy Aven, Carnegie Mellon University
Dr. Hong Shen, Carnegie Mellon University
COL David Beskow, United States Military Academy

Abstract

Before the ubiquity of social media, the information space was dominated by a small number of trusted originators (e.g., news organizations). This paradigm was shattered and replaced with a more diffused information environment where content originates from often unknown actors and is propagated nearly instantaneously. The need to rapidly identify factually inaccurate information (misinformation) in this environment is critical. Though some solutions exist to this problem, most current systems rely on human-driven fact-checking. These systems cannot operate at scale and are subject to significant human bias. In this thesis, I will describe a methodology for a fully automated misinformation detection pipeline that operates on multiple social media platforms. The pipeline leverages natural language processing (NLP) approaches to find, extract, and contextualize claims that could contain misinformation. Another set of NLP models and network science approaches will assign a likelihood to the truth value of these claims to arrive at a final misinformation likelihood score. As part of this thesis, I will validate the methodology in terms of its ability to accurately detect misinformation in labeled datasets as well as its research utility in the social cybersecurity domain by applying it in case studies on multiple social media platforms that focus on diverse topics/communities. The resulting system will serve as a valuable part of the social cybersecurity researchers' toolkit, to be used alongside the BEND framework to characterize the information environment and ultimately inform effective countermeasures.

Contents

1 Introduction

This document describes a proposed thesis that seeks to develop an automated methodology for misinformation detection and test its validity. This system's intended application is social cybersecurity research. Specifically, the ability to automatically identify likely misinformation in large social media corpora would provide a critical and missing analytic lens through which we can characterize the information space and ultimately identify social cybersecurity threats.

The remainder of this paper will justify the need for misinformation detection, position this research in the broader context of the social cybersecurity field, propose a six-part methodology for detecting misinformation, describe how the system will be validated, and lay out the future course of this research.

2 Background and Motivation

2.1 Definitions of Key Terms

Before proceeding further, it is important to establish definitions of some key terms. As social cybersecurity is an evolving field, the common definitions for some of these terms have fluctuated over time, and some do not have clear consensus definitions. For this thesis, we will use the following definitions (sourced from Dictionary.com):

- **Misinformation** - false information spread, regardless of the intent to mislead.
- **Disinformation** - false information deliberately spread to deceive people.
- **Propaganda** - information, ideas, or rumors deliberately spread widely to help or harm a person, group, movement, institution, nation, etc.
- **Conspiracy Theory** - a theory that rejects the standard explanation for an event and instead credits a covert group or organization with carrying out a secret plot.
- **Claim** - an assertion of the verifiable truth of something.

2.2 Research Questions

Using the definitions in the previous section, disinformation is a subset of misinformation that involves intentionally sharing false information to deceive others. The system described in this paper will not attempt to ascribe motive, so we will not attempt to separate misinformation from disinformation. Furthermore, because disinformation is a subset of misinformation, we will avoid using the popular "mis/disinformation" nomenclature common in related work in favor of simply using "misinformation."

This thesis will focus on the following research questions:

RQ1: How can we detect misinformation in large datasets at the post level in a way that is useful for social cybersecurity research? This is the central research question of the thesis. Section 2.3 further motivates its importance, and section 3 provides a study-level breakdown describing how I believe this can be executed.

RQ2: How is misinformation used in information operations? After developing the capability to detect misinformation at scale, I will examine how misinformation has been used (and emerged) in known information operations. Many researchers have answered this question anecdotally by manually identifying misinformation in the past. This study seeks to expand this research capability for future studies.

RQ2.1: Does the base rate of misinformation use differ between the BEND maneuvers? BEND is a framework for describing how people interact in the information space, and it would be useful to study how creating or perpetuating misinformation is used within the overall BEND maneuver set. For example, perhaps negative maneuvers (e.g., Dismiss and Distract) are more likely to contain misinformation than their positive counterparts (e.g., Excite and Engage).

RQ2.2: Do different actors employ misinformation differently in their maneuver sets? Similar to RQ2.1, this question seeks to discover if the use of misinformation is constant across different actors (or types of actors). For example, perhaps political commentators are less likely to use information than medical commentators.

RQ2.3: Is misinformation linked to more effective maneuvers? The previous research questions focus on misinformation in attempted BEND maneuvers, but it is also important to determine if employing or capitalizing on misinformation is associated with more effective information space maneuvers. For example, perhaps Engage maneuvers are more successful when employing sensationalist misinformation than maneuvers of the same type that employ truthful content.

2.3 Importance of Misinformation Detection

Misinformation has become a ubiquitous part of the modern information environment. Bad actors use it to sow confusion in target populations, and it is also an unintentional byproduct of the large and growing information space that dominates the modern internet. Regardless of the intent of those sharing it, unchecked misinformation has the potential to upend democratic elections, subvert public health interventions, and generally create a dysfunctional social environment.

Misinformation has pushed modern society to what some call a ”post-truth” era that is defined by constant suspicion of any information, even provable facts [16]. This suspicion is dangerous as it is nearly impossible for society to function in necessary collective ways. For example, misinformation surrounding vaccines has led to parents refusing to vaccinate their children, leading to outbreaks of deadly, previously eradicated diseases like measles [26].

The World Economic Forum’s 2024 Global Risk Report cites misinformation as the biggest immediate risk to humanity [29]. The report highlights concerns that misinformation could be used to destabilize societies by driving political polarization, potentially resulting in politically motivated violence. Additionally, the report describes how widespread misinformation could have second-order effects, giving authoritarians an excuse to enhance censorship under the guise of protecting people from dangerous misinformation.

Social media platforms publicly recognize the potential risks of misinformation and, in some cases, have attempted interventions to combat it [24]. The platforms are in the best position to address this problem; however, there is limited incentive to identify and address misinformation in any substantive way. Highlighting the amount of misinformation on a platform would be admitting a failure, so they will likely continue to provide cursory bandages that are palatable from a public relations perspective instead of truly addressing misinformation on their platforms.

Given that misinformation will almost certainly continue to be a major part of the information environment, it is important to enable research into ways to mitigate its negative effects. In order to craft effective countermeasures, however, the research community must gain a better understanding of how misinformation is employed and emerges through large-scale studies of information environments. These studies are currently constrained by the absence of a scalable method to detect misinformation.

2.4 Positioning in Misinformation NLP Research

There is an active research community studying misinformation detection, but the current focus of most work differs significantly from this proposed thesis. In particular, many current papers focus on ways to help scale and otherwise enable human fact-checking. The emphasis on human-driven fact-checking is understandable because determining the truth value of a statement requires context and sophisticated understanding, which makes it difficult to automate with existing NLP techniques. Many of these studies seek to ascribe ”check-worthiness” to a claim, which measures how important it is to have a human review a particular claim [7]. Scoring check-worthiness is so popular that it was a multi-year task in one of the most popular research competitions in the NLP space: SemEval [14]. The community made significant progress extracting check-worthy claims. Still, ultimately, these claims are sent to human fact-checkers, which exposes the inevitable problem with this type of fact-checking: it is vulnerable to human bias. The system proposed within this thesis seeks to be completely automated and separated from human decision-making, addressing a key flaw in the popular fact-checking methodologies.

In a few notable studies, researchers have attempted to craft fully automated fact-checking systems [11]. These methods use the same kind of claim detection and extraction methods employed by the check-worthiness studies, but they substitute human reviewers for NLP search techniques and databases of facts. Systems like this are well-designed to identify misinformation in domains with large repositories of immutable facts - but fields with those characteristics are the exception rather than the rule. For example, it might seem tractable to create a database of scientific facts that could be used to check scientific claims. Consider, however, that scientific consensus is ever-shifting, and there is nuance in most scientific fields that does not lend itself to a single set of codified static facts. Moreover, these automated systems are obviously inadequate for evaluating anything pertaining to current events, as it would be impossible to maintain a running database that can keep up with the speed of the modern information environment.

Another popular field that overlaps with this thesis is propaganda detection. As defined in section 2.1, propaganda differs from misinformation in that it does not necessarily have to be false. As a result, unlike the fact-checking field, propaganda detection is not singularly focused on the truth value of claims. Instead, it attempts to recognize signatures of defined tradecraft that signals propaganda. In particular, the NLP-based approaches for propaganda detection are focused on detecting these known techniques and frameworks [27]. There are, however, several network-based propaganda detection approaches that could inform this thesis [17].

2.5 Positioning in Social Cybersecurity Research

BEND is a CASOS-developed framework for characterizing maneuvers in information spaces [3]. It has been adopted by social cybersecurity researchers throughout academia and government. BEND allows researchers to use a common set of terms to classify activities, and it has been used in many studies evaluating real-world information spaces, from the Russian invasion

of Ukraine to Chinese information operations [15, 19]. BEND maneuvers can be automatically detected using the linguistic and network features of social media data using Netmapper and ORA software. This automatic detection enables studies that leverage large amounts of data, providing a nearly instant description of the information space.

BEND studies are most effective when they use other dimensions as lenses through which the information space can be observed in more detail. For example, previous studies use topic modeling, stance detection, and linguistic analysis as overlays to the BEND maneuvers to produce useful analytic results. Studies leveraging BEND and stance detection, for example, can show how users with different beliefs about a subject tend to operate differently in the information space [4]. Similarly, bot detection is a powerful force multiplier when combined with BEND, as it shows not only how bots are being employed in the information space but exactly how they are being applied (e.g., bots are often used to Build communities by artificially inflating an account's network position) [18].

Like bot detection, automated misinformation detection could be a powerful addition to BEND analysis. For example, it is currently unknown if certain BEND maneuvers are more likely to contain misinformation than others (RQ2.1). Answering this question would lend critical insight into how actors operate in the information space. Moreover, studying how different actors use misinformation in their maneuver sets could help describe the playbooks or tradecraft employed by different bad actors (RQ2.2). Finally, while we know that misinformation is employed and emerges in the information space, its effects can be difficult to quantify. Several current studies seek to construct an evaluation framework for BEND effects, which could be used with the misinformation detection pipeline to determine if including misinformation is associated with more successful maneuvers.

3 Misinformation Detection Study Design

The misinformation detection pipeline shown in Figure 1 is the central methodology for this thesis. The goal is to provide a message-level misinformation score and for this methodology to function across the spectrum of social media services.

The orange box shows an overview of the natural language processing pipeline. The process starts by extracting claims from social media text, clustering similar claims, and finally performing linguistic analysis to detect logical fallacies and find claims at odds with each other. The blue box shows an overview of the network science pipeline where social media is represented as various networks that can be used to help identify the likely presence of misinformation and provide information about source reliability. Combining these pipelines results in a final misinformation likelihood score.

The following sections will describe each of the proposed studies (shown in red in Figure 1) to explain their motivation, expected outputs, and feasibility.

3.1 Data

Many of the studies listed in the following section rely on supervised modeling techniques that require high-quality, labeled data. Moreover, even the unsupervised aspects of the pipeline will need to be validated using (smaller) sets of labeled data. The goal is to create datasets that are large enough for these studies but not so large that creating them becomes the main effort of this thesis.

A major requirement of this thesis is to be platform-agnostic. It is impossible to include every social media service in the dataset, so the goal is to include platforms that represent the major

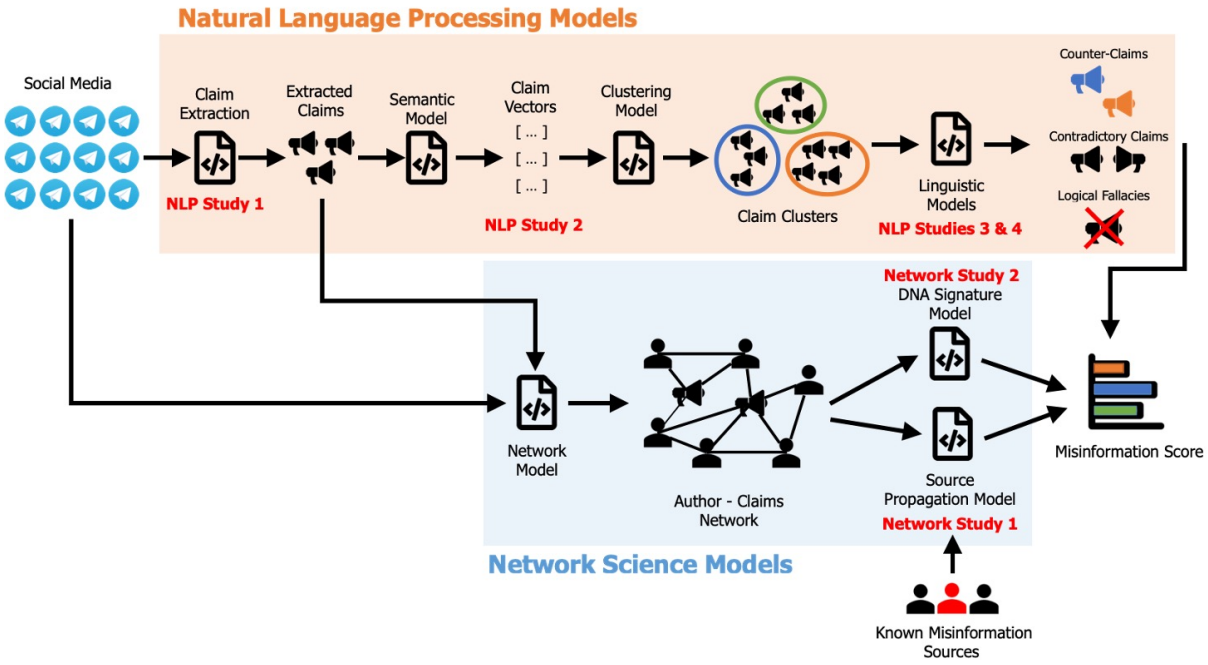

Figure 1: An overview of the misinformation detection pipeline. The orange box surrounds the NLP studies, and the blue box surrounds the network studies. Red text corresponds to specific studies that are proposed in detail in this section.

types of social media. I will focus on Twitter, Reddit, Telegram, and Facebook. Twitter is distinct from the other platforms in that it imposes a character-length restriction and has many features that are not present on other platforms (e.g., hashtags as global topic tags). However, there are many Twitter "clones" such as Mastodon and BlueSky that could leverage models created using only Twitter data. Beyond Twitter-like social media, there are many forum-based social media services that are similar to Reddit. The platforms are divided into user-specific sub-communities that are typically topic-oriented. Finally, there are some social media platforms that come from peer-to-peer messaging services that are comprised of large public-facing community groups that fracture into smaller public and private groups. This study will use Telegram and Facebook to represent this type of data.

Aside from being applicable to different social media platforms, this thesis also aims to be topic-agnostic. I will focus on three diverse topics: COVID-19, U.S. elections, and the Russian invasion of Ukraine. These topics were selected because they cover key subject areas broadly important in the information environment: medicine/public health, politics, and war. Additionally, these three topics were selected because there are existing Twitter and Reddit datasets on each of these topics. Both platforms significantly restricted access to their data in 2023, so acquiring data for more recent events is cost-prohibitive (or impossible).

Figure 2 shows how these datasets will be combined into a master dataset to facilitate the NLP studies. Each data subset contains 2,000 messages, creating a master dataset of 24,000 observations. This data size was chosen to be consistent with the size of the datasets used in CheckThat! claim detection competitions [7]. Fortunately, most of this data already exists in CASOS-owned datasets, but I will need to collect a small number of Telegram data on U.S. Politics, as well as the full Facebook dataset. This data is being procured from a vendor to support multiple CASOS studies and will be available by summer 2024.

Platform	Topic	Data Size	Have Data?	Claims Tagged
Twitter	COVID-19	2,000	Yes	Yes
	US Politics	2,000	Yes	No
	Russian Invasion	2,000	Yes	No
Reddit	COVID-19	2,000	Yes	No
	US Politics	2,000	Yes	No
	Russian Invasion	2,000	Yes	No
Telegram	COVID-19	2,000	Yes	No
	US Politics	2,000	No	No
	Russian Invasion	2,000	Yes	No
Facebook	COVID-19	2,000	No	No
	US Politics	2,000	No	No
	Russian Invasion	2,000	No	No
	Total	24,000		
	Total to Label	22,000		

Figure 2: A description of the data that will be used to conduct the NLP studies.

While most of the data is already on hand, the data must be labeled to be useful in the NLP studies. Specifically, three different human labelers will tag each of the social media messages as containing a claim or not, and if the message contains a claim, the labeler will highlight it in the message. 2,000 of the messages were previously tagged as part of a CheckThat! competition dataset [7], but the remaining 22,000 claims need to be tagged. Because there are three reviewers, I cannot use the commonly applied Cohen's Kappa to assess interrater reliability, so I will instead use Fleiss' Kappa, with the goal of reaching at least 0.80.

Using only the extracted claims from this dataset, I will create another set of labels denoting if a logical fallacy was used. The specific number of claims I can label in this step will depend on the number of claims extracted from the original messages. The goal will be to produce the largest balanced dataset possible. I will also label at least 1,000 claim pairs per platform as similar or not similar (see section 3.2.4); and contradictory, counter, or neither(see section 3.2.4).

In addition to this first dataset, I will also need to develop a second dataset that can be used for the two network studies in the detection pipeline as well as the two validation studies. This dataset is described in Figure 3. The sources for this dataset will be the same as those shown in Figure 2], but the labeling process will be substantially different and, notably, more time-intensive. As a result, this dataset will be much smaller than the first, with only 3,000 total messages.

For each sub-dataset, human labelers will tag 250 claims as containing misinformation or not. This will be a major effort, as the human-labelers will have to conduct research on each claim to accurately verify its truth value. The labelers will be given instructions on how to use existing fact-checking repositories to aid this process. Unlike the previous dataset, we will not accept imperfect interrater reliability and will instead require consensus from all three labelers before labeling a post as misinformation. Human bias is a real concern for human-tagged misinformation, and the consensus-based approach is designed to limit these biases. The final dataset will incorporate these tagged claims back into the original social media messages and will be combined with other messages that did not contain claims.

Overall, the data collection process is trivial, but the labeling effort will be substantial. Fortunately, these labeled datasets will be useful to future researchers in the same domain and likely in many other social cybersecurity research efforts in the lab. To the extent possible, I plan to

Platform	Topic	Data Size (Claims)	Have Data?	Misinformation Tagged
Twitter	COVID-19	250	Yes	No
	US Politics	250	Yes	No
	Russian Invasion	250	Yes	No
Reddit	COVID-19	250	Yes	No
	US Politics	250	Yes	No
	Russian Invasion	250	Yes	No
Telegram	COVID-19	250	Yes	No
	US Politics	250	No	No
	Russian Invasion	250	Yes	No
Facebook	COVID-19	250	No	No
	US Politics	250	No	No
	Russian Invasion	250	No	No
	Total	3,000		

Figure 3: A description of the data that will be used to conduct the network studies in the detection pipeline and the validation studies.

publish as many of these datasets as possible in venues that will support future research. I anticipate this (painful) data-curation process will ultimately be a major contribution of this thesis.

3.2 Proposed Studies

3.2.1 NLP 1: Claim Detection and Extraction

Detecting and extracting claims from social media text is the cornerstone of this entire research effort. Every other proposed study uses these claims downstream, so it is critical that this first method performs well.

As defined in section 2.1, a claim for the purposes of this research is an assertion of the verifiable truth of something. Unpacking this definition means that a claim is differentiated from other common types of statements by the fact that it can be (theoretically) verified as either true or false. For example, a claim about the COVID-19 vaccine could be, ”The COVID-19 vaccine is effective at reducing illness severity,” while a similar opinion could be, ”I think everyone should take the COVID-19 vaccine.” Interestingly, the opinion has a truth value - it is either true or false that the person has this particular belief - but it is not possible to verify. This study will focus on differentiating between claims and other types of statements using their unique linguistic features.

Identifying and extracting claims is a task that is well-suited for modern NLP. The remainder of this section will provide a short overview of the current state of claim detection research that highlights existing gaps relevant to this thesis, propose a two-part study for addressing those gaps, and conclude with a discussion of the data that will be used.

Existing Work Claim detection has emerged as a fairly common task in recent NLP studies. Interestingly, however, few studies focus on social media text, which is considerably different from things like news and scientific journal articles (two common sources of text for NLP studies). An exception to this is two popular NLP academic competitions that have tackled claim detection in the past several years: SemEval and CheckThat! [7, 14]. Both of these competitions exclusively used Twitter data, and both of them narrowly focused on specific domains. SemEval-2023 provided Tweets with medical claims, and CheckThat! 2019 and 2021 provided Tweets about COVID-19 and U.S. politics.

Beyond focusing on a limited domain and data source, the SemEval and CheckThat! tasks have both shifted to focus on check-worthiness. This has influenced NLP papers outside the

competitions to focus on the same [8,23]. Check-worthiness scores not only whether a statement is a verifiable claim, but also how important it is to fact-check the claim. This is an important concept in the field of semi-automated fact-checking because it allows a system to prioritize claims before they are sent to human fact-checkers. The movement towards check-worthiness may seem like a subtle shift from claim detection, but it has essentially reframed the task to be about categorizing known claims by their importance. This task modification makes this current (popular) brand of claim research inapplicable to this thesis.

Proposed Work To facilitate the goal of finding claims in social media, the methodology proposed in this thesis is divided into two sections: binary message classification and span extraction. The general idea is to use a classification model to identify social media messages that contain a claim, then use a span extraction model to isolate the claims from the rest of the content in each message.

It is possible that the span extraction step could function without the need for message classification, but the study plan is to perform a two-stage modeling approach to ensure optimal performance. A perfect span extraction model would be able to take input without a claim and simply output nothing. In reality, the model will likely have imperfect precision (i.e., it will sometimes label spans as claims when they are not), so this study will use the message classification model to filter out messages that do not contain a claim. In practice, this will allow us to train the span extraction model using data with known claims, which should result in better performance as the model's task is more specific.

Classification The first phase of the study will involve creating a model for binary classification that labels messages as containing a claim or not. Several machine learning approaches could be used to build this classifier, but transformer-based encoder models (e.g., BERT) are particularly well-suited to the task. These models are pre-trained on large corpora of English text, performing a masked language task. Masked language modeling hides a random proportion (typically 15%) of the tokens in a document from a model and requires the model to fill in the gaps. Previous research has demonstrated that task-tuning base models pre-trained on masked language is an effective method for classification [2].

This study will focus on the following pre-trained base models: BERT, distilBERT, RoBERTa, and distilRoBERTa. BERT and RoBERTa represent slightly different transformer model design choices, while the models with "distil" are smaller versions of the full models. It is not always obvious why a specific transformer model design results in better performance on certain tasks. Hence, this study follows the common practice in modern NLP studies to test several model architectures.

For each base model, I will perform task-tuning over multiple sets of hyperparameters, evaluating model performance on a validation set (the specific data structure is discussed in the Data section below). There are many hyperparameters that can be modified when fine-tuning these models. I will focus on three that have been shown to be most important in previous classification studies: batch size, learning rate, and number of epochs. The batch size determines how many samples are processed before the model updates, learning rate determines the extremity with which the model weights are updated, and the number of epochs determines the overall number times the model processes the full training set. Previous work has demonstrated a general range for these hyperparameters: batch sizes 16 and 32 (2 levels); learning rates between 2e-5 to 5e-5 (4 levels); and number of epochs 2 to 5 (4 levels). The full experimental design will fine-tune 128 models for evaluation. The full experimental design is shown in Figure 4

Models	Levels	Values Used
Transformer-based Encoder Models	4	BERT, RoBERTa, distilBERT, distilRoBERTa
Hyperparameters	**Levels**	**Values Used**
Batch size	2	16, 32
Learning rate	4	2e-5, 3e-5, 4e-5, 5e-5
Number of epochs	4	2, 3, 4, 5
Total Experiments	**128**	

Figure 4: An experimental design for testing different models and hyperparameters for claim extraction.

Beyond fine-tuning for classification, I will also study the effects of domain-adaptive pre-training (DAPT) [12]. To avoid training a large number of models, I will perform DAPT only on the two best models from the first fine-tuning experiment. DAPT is done before the task-tuning step and is done by performing additional masked language pre-training on a set of text data that is relevant to a specific domain. In this case, I will use a balanced set of social media text from Twitter, Reddit, Telegram, and Facebook with the goal of pre-training the model to work well with social media in general. I will also create models using data from each social media platform in isolation. Ultimately, I will train five DAPT models: Twitter, Reddit, Telegram, Facebook, and combined social media - all for the top two model candidates from the initial fine-tuning experiments, resulting in 8 additional models.

Sequence Tagging After identifying social media posts that contain claims, a sequence tagging model will be used to isolate the claim from the rest of the text. In short-text social media such as Twitter, this task will be easier as the content limit does not allow for much extraneous text. On Reddit and Telegram, however, longer posts are more common and pose more of a challenge. This study aims to develop a single extraction model that is performant on data from all three platforms, but this may need to be expanded to include short-text and long-text versions of the model.

Because of the flexibility of the transformer modeling framework, sequence tagging can leverage the same experimental design proposed in the previous section (Figure 4). Sequence tagging can be framed as a binary classification task where the text is broken into sequences, and the model classifies each sequence as containing a full claim or not. The overall study design will use the same hyperparameter ranges as the classification study. Furthermore, because of the similarity between this task and the previously described classification task, we can reuse the same DAPT models without the need for repeated computation.

Some recent work has suggested that decoder-based LLMs (e.g., GPT) can perform sequence tagging without transforming the problem into a classification task [6]. However, I find it unlikely that this type of LLM will be able to perform this task with a single prompt in a sufficiently generalizable way. To test this hypothesis, I will experiment with zero-shot and few-shot learning using these LLMs. This will provide a useful comparison for the encoder-based sequence tagging approaches.

Data The data for both studies will come from the process detailed in section 3.1 with some important model-specific pre-processing. For the classification models, the data will include raw social media text with a binary label that corresponds to whether or not the text contains a claim. For the sequence tagging model, the data will need to be processed into sequences (sentences in this case) with labels denoting whether the sequence contains a complete claim.

Both datasets will need to be split into train and test sets using a stratified sampling approach that maintains the same proportion of positive/negative labels as well as social media platform sources (e.g., train and test should contain the same proportions of Twitter, Reddit, Telegram, and Facebook).

3.2.2 NLP 2: Claim Clustering

After extracting claims, the next step is grouping similar claims. This is an important step in the process because it will determine which claims need to be compared by the downstream linguistic models. This study will use two methods to cluster claims: one based on the popular BERTopic framework and another that leverages entity graphs.

BERTopic's popularity has grown over the past few years, and it is often treated as the gold standard for clustering text (especially short-text) documents. While powerful, this method is limited in that it is only able to group texts that are semantically similar - which might not be universally true for similar claims. To combat this issue, I propose adding a second layer to the claim clustering pipeline that leverages named entity recognition (NER) to create entity graphs that provide a different type of similarity entirely. Depending on their performance, these two approaches could be used together or separately in the final pipeline. If used together, the models will be used in a multi-model voting strategy that places the highest likelihood of correct clustering in cases where the models agree.

Existing Work The first proposed method for clustering claims is based on the BERTopic framework that has largely replaced traditional algorithms like LDA and LSA for small-text topic modeling. BERTopic is not a set algorithm but a general framework for clustering text documents based on their semantic meanings [9]. The backbone of BERTopic (and the reason for its name) is the use of encoder-based transformer models to translate unprocessed text into a high-dimension vector representation (a process commonly called embedding). The encoder models that create these embeddings are trained with pairs of semantically similar sentences such that the embeddings produced by the model place similar chunks of text near each other in this high-dimension space. Unlike the study described in 3.2.1, these transformer models are already tuned to the task at hand and do not need further fine-tuning (typically).

After achieving an appropriate semantic embedding, any clustering algorithm could be used to find similar sentences. In practice, however, clustering algorithms tend to perform poorly in high-dimensional spaces, so the convention is to perform dimensionality reduction before clustering. Finally, there are several techniques that label the resulting clusters by looking at word frequencies, term frequency/inverse document frequency (TF/IDF), and even using LLMs to assign topic labels.

In addition to leveraging the BERTopic framework, this study will also attempt a less-used methodology that combines named entity recognition (NER) with bipartite network clustering to cluster similar documents according to the entities they mention. The SciClops framework performs a similar method for entity resolution, but it is only tuned and evaluated on scientific claims in a narrow band of topics [21].

Proposed Work

Semantic Clustering As mentioned above, the semantic clustering methodology used in this paper will be guided by BERTopic. However, it is not sufficient to call this study simply an

application of BERTopic, which is really just a framework for clustering text embedded in semantic space. Many different algorithms and modifications of those algorithms can be used at every step of the BERTopic process. This study will attempt to identify the combination of models and hyperparameters that work best for clustering claims across social media platforms and topics.

The first step in this study is to identify an embedding model that functions well with claims. Common embedding models are typically applied without much consideration or evaluation. This study will test several embedding models that are commonly on the top of community leaderboards: Microsoft's MiniLM and mnet models as well as a distilRoBERTa variant fine-tuned to this task. The evaluation strategy will take a random sample of 100 documents from each major topic area (COVID-19, U.S. elections, and Russian invasion of Ukraine) from each social media platform (Twitter, Reddit, Telegram, and Facebook). Using either cosine similarity or Euclidean distance (depending on which metric was used to generate the embedding model in question), I will perform pairwise comparisons between each of the 300 documents (44,850 total comparisons). The best embedding strategy will provide the greatest difference between the average distance to embeddings outside the same topic and the average distance to embedding within the same topic. This process will be repeated 100 times with different random samples (for a total of 4,485,000) comparisons.

With the best embedding model selected, the next step is to experiment with different combinations of dimensionality reduction and clustering frameworks. In previous work, I have found that the combination of UMAP for dimensionality reduction and K-Means (with the k-parameter tuned with the elbow method) to be the best combination, but I will also experiment with PCA for dimensionality reduction and HDBSCAN for clustering. The performance of these clustering algorithms will be evaluated using a subset of manually tagged data (detailed in the Data section below).

Entity Graph Clustering The BERTopic pipeline is an effective and proven way to cluster semantically similar text, but its performance relies on the presence of semantic similarity. It is likely that at least some claims that are about the same thing are not semantically similar - especially if these claims run counter to each other (which is a focus area of this thesis). In these cases, however, it is still likely that these claims contain references to the same named entities. The method described in this section leverages these common entities instead of semantics to cluster similar claims.

The first step in this study is to extract named entities from claims. This is a well-studied subfield within NLP, and there are many different models that could be used for this task. I will evaluate a small encoder model tuned for named entity recognition (NER) based on distilBERT, a larger encoder model based on RoBERTa, and a version of a popular encoder-decoder model called T5. These models will be evaluated against a small subset of human-tagged data. The goal of this study is not to reinvent the benchmarks for NER, but to instead confirm that these models are performing as expected in this use case.

After extracting the entities from each claim, I will create a bipartite network with edges between a claim and each of the entities mentioned in that claim. This bipartite network can then be used to cluster similar claims based on their shared entities. Specifically, I will fold the claim x entity network (multiplying it by its own transpose) to create a claim x claim network where the edge weight corresponds to the number of shared entities in the bipartite graph. I will then use Louvain clustering (and possibly other methods, depending on the results) to cluster the claims. I expect this approach will be especially effective at differentiating claims that are about the same topic (e.g., COVID-19) but are about different specifics (e.g., government interventions

vs/. viral origins).

Data The data used in these studies will be the same labeled claims described in section 3.1 and used in the first NLP study. Unlike the claim detection work, neither method proposed here is supervised, so there is no need to tag a huge amount of data for training and test purposes. Instead, I will label a smaller subset of the data, binning social media messages into topic categories that span all three social media platforms. For example, I will manually identify Tweets, Reddit posts, and Telegram messages that all discuss COVID-19 masks. These labeled sets will be used to validate the unsupervised clustering methods.

3.2.3 NLP 3: Logical Fallacy Detection

After identifying claims, we can conduct the first study aimed at detecting the indicators of misinformation directly by identifying logical fallacies. It is undoubtedly true that logical fallacies are used regularly when making factually true claims due to the arguer's linguistic or logical sloppiness. The use of logical fallacies is, however, an indicator that the information contained in a claim could be false. This study seeks to build a classifier that can detect logical fallacies across social media platforms and topic areas.

Existing Work Logical fallacy detection is an NLP subtask that dates back to early computational linguistic studies. Interest in this subproblem can probably be attributed, to some degree, to the strong presence of researchers trained in traditional linguistics who have become interested in computational linguistics. As a result, there are many methods that span the timeline of popular NLP approaches that seek to detect logical fallacies in text. And, as with most other areas of NLP, recent studies have found that transformer models are particularly adept at this NLP task [13, 22].

While these transformer-based papers show the potential power of these models for detecting logical fallacies, they frame the problem as a multi-classification problem where the goal is to train a model that is able to differentiate logical fallacies. In some cases, the input data to the models always contains a logical fallacy, and the task is to label it with a specific fallacy. These findings are encouraging for this research, but the existing models have entirely different goals than my proposed work.

Proposed Work Unlike existing papers that seek to classify specific types of logical fallacies in a multi-classification task, this paper will instead reframe the task to binary classification. The objective is to train a model that can flag that a logical fallacy is present in a claim, but we are not concerned with the specific fallacy. This task simplification should provide significant performance gains over current methods. Furthermore, given that the multi-classification version of this has proven to be tractable, simplifying the task should increase the likelihood that this study will be successful.

This study will evaluate three different classes of transformer models: encoder-only (e.g., BERT), encoder-decoder (e.g., T5), and decoder-only (e.g., GPT). Unlike the previous NLP studies in this proposal, there is no significant existing work attempting binary classification for the presence of logical fallacies. As a result, I am less confident that a specific type of model will be the most performant (though I strongly suspect encoder-only models will work well given their application in other classification tasks). So, this study will apply a similar methodology to the one described in 3.2.1 to train and test four different encoder-only transformer models: BERT, distilBERT, RoBERTa, and distilRoBERTa. Again, I will use the set of hyperparameters

outlined in this initial study. Additionally, I will again attempt to perform DAPT to improve the two best models, but in this case, I will use data that contains only claims to perform the DAPT (instead of different types of social media data).

After testing the encoder-only models, I will test at least one encoder-decoder model (Google's T5 model). The T5 model was trained on a huge number of different tasks, and it is capable of performing classification tasks without additional fine-tuning (i.e., a zero-shot approach). I expect the zero-shot method will yield inferior results, so I will also attempt to fine-tune a T5 model (likely a smaller variant than the full-sized model) using labeled data.

For a final comparison, I will attempt zero-shot and few-shot classification using a decoder-only LLM. I will likely use GPT-4, which is currently the gold standard for commercially available decoder-only models, but I will also experiment with more novel models, such as MistralAI's new Mixtral-8x7B model. The zero-shot approach will involve some prompt engineering but will otherwise present the model with a claim and request a binary label corresponding to whether it contains a logical fallacy. The few-shot approach will provide a set of positive and negative samples with labels along with the prompt.

It is my strong preference not to use decoder-only LLMs. These models often provide useful results on subsets of data, but fail to generalize. However, given their popularity, especially in the field of computational social science where I anticipate this work being most relevant, it is worthwhile to test some decoder-only LLMs if for no reason other than to provide a comparison for the other modeling approaches.

Data All of the modeling approaches proposed in this study will rely on labeled data. Specifically, I will need to label a large set of claims as either containing a logical fallacy or not. The biggest obstacle to creating this data is finding a sufficient amount of claims that contain logical fallacies. To combat this issue, it may be useful to leverage decoder-only LLMs to generate synthetic claims with logical fallacies. While I am skeptical that these models are the best choice for classification tasks, they are undoubtedly well-suited (and, in fact, were designed) to generate natural language.

In the previous NLP studies, it was important to ensure that the labeled data contained a balanced set of messages from different social media sources. Given that this model operates on claims (and no other text features from the original source message), this is less of an issue in this study. It will, however, be important to balance the topics of the claims between COVID-19, U.S. elections, and the Russian invasion of Ukraine to avoid training a model that does not generalize to other topics.

3.2.4 NLP 4: Counter and Contradictory Claims Detection

The final NLP study in this proposal seeks to detect another indicator of misinformation based on claims and their counters. Specifically, this study attempts to match claims with a claim that either directly contradicts or runs counter to the original claim by directly comparing the claim clusters created in the previously described study. For the purposes of this thesis, a contradictory claim is a claim that directly refutes another (e.g., the COVID-19 vaccine is healthy vs. the COVID-19 vaccine is unhealthy). A counterclaim is a claim that does not directly refute the claims of the first but instead presents another claim that is meant to discount the original (e.g., the COVID-19 vaccine is healthy vs. the government's COVID-19 policies are only meant to protect the economy). Detecting counter and contradictory claims can indicate the presence of misinformation because these claims are typically controversial and, thus, elicit a response from those who disagree.

To detect counter and contradictory claims, this study will use a two step methodology that uses both classification models as well as semantic networks. I expect that using both methods together with a voting strategy will yield the best performance, but it is possible that using only one of the methods will be sufficient.

Existing Work As with the other NLP studies proposed in this paper, recent work in this domain has demonstrated that transformer models are well-suited to detecting contradictory claims [20]. In this case, however, even the most recent studies that have tackled this task are somewhat out of date and use older transformer models. The success of these past studies is reassuring that I can create an improved model by simply updating these approaches with the current slate of transformer models and fine-tuning strategies.

Existing research also demonstrates how semantic networks can be used to detect contradictions in text data. Many of these studies are rooted in traditional linguistics, and, while they prove that this method has conceptual merit, they are not designed to be used in a practical, scalable methodology. One notable study designed a single model for claim extraction and comparison, leveraging semantic networks in their pipeline; however, the model is largely focused on claim detection and is ultimately bogged down by attempting to perform several NLP tasks in a single model [5].

Proposed Work

Classification Model The classification modeling approach in this study is very similar to the one proposed in section 3.2.3 for logical fallacy classification. In this case, I will label a pair of claims as either containing a counterclaim, a contradictory claim, or neither. Notably, this differs from the other classification models proposed in this paper because this approach is not a binary classification task and is instead a categorization task. It would be possible to formulate this as a binary classification task that identifies a pair of claims as containing either a counter or contradictory claim, but there is potential downstream value of classifying these types of claim rebuttal separately. Specifically, it is possible that, say, the presence of a directly contradictory claim is more closely associated with the presence of misinformation when compared to a counterclaim.

As with the logical fallacy detection models, there is limited research in this specific domain, so I propose testing encoder-only models, encoder-decoder models, and decoder-only LLMs. Shifting to a general classification problem from binary classification does not require meaningful modification to the model training and test methodology. The key difference between these studies is the data preparation, described in detail in the data section below.

Semantic Networks To enhance the classification model's predictive ability, I will also attempt to create a model that uses semantic networks to identify pairs of counter and contradictory claims. Semantic networks process natural text into a network structure that codes concepts and entities as nodes and relationships between those concepts as edges. Contradictory claims should contain similar concepts but opposite relationships between them. I am unaware of previous work that applies the same analysis to counterclaims, so the next part of this study will use labeled pairs of counterclaims to identify the relationships between their entity graphs. The ultimate goal is to find a set of features present in these subnetwork pairs that differentiates these pairs from unrelated claims.

Data Both the classification and semantic network models will require labeled input/validation data. For this study, the data will be comprised of pairs of claims and a category: counterclaim, contradictory claim, or neither. The data to create these labeled pairs will come from the claim clustering study proposed in section 3.2.2. The goal is for these models to be able to differentiate the claim pairs that are within the clusters created by this earlier study, as that is their intended use. These pre-clustered claims will be focused on the same general topics, which will allow these models to focus on detecting the linguistics associated with contradiction and countering instead of simply identifying topics.

3.2.5 Network 1: Unreliable Source Propagation

The goal of the aforementioned NLP studies is to identify indicators of potential misinformation at the message level. This study shifts the focus to the network structure of social media data and is the main method used in this study to provide insight into the likely truth value of a claim. Unlike human-driven fact-checking systems that provide hard ”true” or ”false” labels to claims, this method instead outputs a score that is based on the probability of a claim's truth value based on the author's relationship to unreliable information sources via their social network. The methodology is split into two main steps: labeling external data sources (i.e., URLs) according to their propensity to contain false information and propagating the derived source reliability through a social network.

Existing Work Many existing repositories score websites based on their propensity to share false information (e.g., Media Bias Ratings) [1]. To augment these existing repositories, several recent studies propose methods to label the reliability of URLs based on features that can be accessed through basic web scraping methods [28]. In general, identifying the trustworthiness of online sources is a well-studied research area, and the current best efforts will be sufficient for use in this methodology.

Similarly, this study will leverage current research into trust propagation through social networks, another well-studied field. Current work suggests several different ways to perform this propagation [10, 25]. Notably, however, the majority of the available studies that perform trust propagation on social media use only Twitter data [30]. While many of Twitter's network characteristics are ubiquitous in nearly all social media platforms (e.g., the ability to reply to another user's message), Twitter has some unique functionality that is not mirrored on other platforms. For example, there is no perfect analogy for a ”retweet” on Reddit or Telegram, and Telegram and Reddit do not use hashtags as a global tagging mechanism. The methodology in this study will need to be robust to various social media platforms - which may mean developing platform-specific algorithms if a one-size-fits-all approach does not perform well.

Proposed Work The first part of this study will extract all of the URLs from a social media corpus and assign each one a reliability score. As mentioned above, this study will largely implement existing methods for labeling URLs' reliability and will rely heavily on existing databases that tag websites' reliability. If there are many URLs that do not exist in the existing repositories, I will explore on-the-fly tagging techniques that rely on webscraping. I will make every effort possible to avoid including webscraping as part of the final pipeline because it is a time consuming and unreliable process - but the option remains in case the existing databases are not complete enough for our task. The final reliability score for each URL will scale from 0 to 1, where 1 corresponds to a completely reliable source.

With the URLs labeled with reliability scores, I will assign an initial score to each user based on the average reliability of the URLs they are directly associated with. On Reddit, this will only include URLs that a user directly posted. On Twitter, it may be useful also to include URLs that appear in retweets. Finally, on Telegram, URLs that appear in post forwards will also be used.

The next step(s) will be to propagate the initial user-level reliability scores using social connections between the users. This process will lean heavily on existing trust propagation methods, as described in the previous section. The propagation step is important, as some people do not post many URLs on social media. In these cases, we need to infer a reliability score for these users based on their social position among their peers. For example, a user who posts no unreliable URLs but is close friends with ten other users who regularly post unreliable information should be tagged as potentially unreliable as well.

Data and Validation This study will use Twitter, Reddit, Telegram, and Facebook data on COVID-19, U.S. elections, and the Russian invasion of Ukraine, as described in section 3.1. Unlike the NLP studies, however, this work will use the data described in Figure 3, which contains human-assigned tags denoting the presence of misinformation.

Using this dataset will provide the opportunity to validate the model using the following process. First, I will group the authors into three groups: reliable, unreliable, and unknown reliability (based on reliability scores calculated by the model). I will then calculate each author's propensity to share misinformation by dividing the number of misinformation messages by their total shared messages. I will then compare the authors by reliability group, identifying statistically significant differences in the propensity to share misinformation. If the methodology functions correctly, the more reliable authors should have a lower propensity to share misinformation than those with lower reliability scores.

To do the group comparisons, I will use a pairwise permutation hypothesis test that will test the mean misinformation ratio for each group against the others. I also plan to run an ANOVA test for completeness (which can test all three groups in a single test), but I strongly prefer the probabilistic interpretation and lack of frequentist assumptions afforded by pairwise permutation hypothesis testing.

3.2.6 Network 2: Dynamic Network Signature

The final study in the misinformation detection pipeline will use dynamic network analysis to identify misinformation signatures in social media datasets. The hypothesis behind this study is that there are detectable differences between social networks where misinformation is propagated and those where it is not. Ultimately, the goal of the study will be to identify users who are sharing likely misinformation and to identify claims that appear in these misinformation-supporting networks. Thus, these results can feed into the message-level misinformation score through the author (source) as well as the claim contained in the message. Fitting with the overall theme of this methodology, this method will certainly not provide a definitive misinformation label. Still, it will be useful in determining the likelihood that misinformation is being spread among groups of users and using certain claims.

Existing Work There is significant existing research that describes the network characteristics of communities that share misinformation, but these studies do not compare these characteristics to those of similar communities that share legitimate information [25]. These studies are not typically designed to identify misinformation in large datasets but rather focus on describing the ways that misinformation is propagated with the ultimate goal of crafting interventions. While

these studies do not demonstrate whether the network characteristics of misinformation-sharing communities are distinct from communities that do not share misinformation, they do show that misinformation-sharing networks have some common features. This is encouraging that it may be possible to use these features to differentiate misinformation networks (which is the goal of this study).

As with the other network-based study proposed in section 3.2.5, the existing work that is relevant to this study is almost exclusively focused on Twitter data. Again, this study will attempt to be generalizable to different types of social media, so it will be developed using Twitter, Reddit, Telegram, and Facebook data. The goal is to develop a universal methodology, but I may need to explore platform-specific modifications to boost performance.

Proposed Work Unlike all of the other studies in this proposal, this method will be entirely based on a case-study approach. Specifically, I will identify subgraphs within large social media datasets that contain significant misinformation spread and similar subgraphs that do not contain misinformation. Next, I will generate network statistics for these subgraphs using various ORA reports (which generate a huge number of network statistics). Finally, I will group the network features for the misinformation groups and the controls, generating distributions of these features, and perform statistical analysis to determine if there are statistically significant differences in the subnetwork features.

I anticipate primarily using permutation testing to compare the distributions of network features of misinformation subgraphs and the controls. These methods do not rely on frequentist statistical assumptions, instead leveraging simulation and resampling approaches to provide probabilities of statistical differences (instead of the often-misrepresented p-values produced by t-testing and other frequentist methods).

Beyond comparing individual statistics with permutation testing, I will also attempt to fit a decision tree model to differentiate between misinformation networks and the controls. Importantly, I will not use methods like random forest, even though they typically outperform basic decision trees, because the purpose of this modeling effort is descriptive. I am seeking a human-readable description of the statistics that differentiate subgraphs that contain disinformation from those that do not.

Data This study will use the same datasets described in section 3.1 but will focus primarily on the network features of the data. I will also need to label subgraphs in the network that contain misinformation. This will be fairly straightforward as all of the proposed topic areas that are represented in the data (COVID-19, U.S. elections, and the Russian invasion of Ukraine) contain many previously identified misinformation narratives. The more difficult task will be identifying the control subnetworks that do not contain misinformation but are otherwise similar to the misinformation networks. For example, subgraphs in the COVID-19 data that discuss the misinformation narrative that the COVID-19 vaccine contains a tracking device should be compared to subgraphs that are also skeptical of the COVID-19 vaccine for reasons that do not contain misinformation (e.g., concerns that the vaccine was rushed).

4 Detection Model Validation

After developing the misinformation pipeline, it will be important to validate it, both in terms of its ability to find misinformation and its research utility. While each sub-study proposed in section 3 includes internal validity checks, the work proposed in this section is meant to evaluate

the pipeline as a whole, answering the key question of how well this detection strategy actually works. The process used to create the data for the validation studies is described in section 3.1 and shown in Figure 3. The remainder of this section will describe each validation study in detail.

4.1 Misinformation Detection Pipeline Validation

The first validation study will examine how well the fully implemented detection pipeline identifies misinformation in social media datasets. As such, the first task will be to process the datasets using the pipeline, outputting the misinformation scores for each message. The misinformation score is a continuous measure of a message's likelihood of containing misinformation, but it will be important to convert this score into a hard label for validation purposes. To determine the optimal score threshold that should be used to label a message as misinformation, I will try a range of possible thresholds and calculate the F1 statistic for each (F1 can be computed because we have the human-assigned labels to act as ground truth). This will construct a curve showing how F1 changes across the range of thresholds, and we will select the threshold with the highest corresponding F1 statistic. The statistic score calculated through this process will provide a measure of the pipeline's ability to identify misinformation.

Because I am using four platforms and three topic areas, it will be important to conduct an error analysis to determine if the pipeline fails to generalize across either feature. Ideally, it would be useful to employ tests of statistical significance to discover if different topics or platforms are associated with lower F1 statistics, but the overall number of data points (12 total combinations of topic/platform without replication) is too low for this kind of analysis. Instead, we will look for dramatic decreases in F1 in any of the topic/platform combinations.

Finally, this study will use a leave-one-out approach to assess the relative importance of each sub-model in the pipeline. This process will iteratively remove one sub-model (e.g., logical fallacy detection) and recompute the F1 statistic for all topic/platform combinations. The most important sub-models will be associated with the biggest degradation in F1. Applying the same error analysis technique described in the previous paragraph, I will examine whether sub-model performance is consistent across platform and topic.

4.2 Reserach Utility Validation

After validating the detection pipeline's ability to detect misinformation, the final validation study will examine the pipeline's utility in research. In particular, this study will examine how similar the conclusions a researcher would draw from a study using model-generated labels are to a hypothetical study conducted using ground truth labels. It is impossible to anticipate and validate all potential uses for a disinformation pipeline, so this study will pick a single expected research use case and evaluate how well the detection pipeline performs. Specifically, this study will evaluate how well the model identifies the proportion of misinformation scores used within each BEND maneuver. Identifying how misinformation is used and/or emerges within the context of the BEND maneuvers is a key element of this thesis, and this specific research use case will demonstrate the system's efficacy in this task.

This validation study will use the same data as the previous one, and I will also retain the misinformation labels generated by the first study. Next, I will use ORA's BEND report to generate labels at the message level, identifying the BEND maneuvers contained in each message. The model-defined misinformation labels and BEND maneuver labels will be combined to create a rank-ordering of the BEND maneuvers that employ the highest proportion of mis-

information. I will repeat the process of creating a rank-ordering of BEND maneuvers using the human-assigned misinformation tags. The degree to which the rank-ordered list created using the model-defined labels is similar to the list created using the ground truth labels will be a measure of the pipeline's research utility.

Ideally, the model and ground truth data would lead to identical results, but this is unlikely in every case. Because the results of the experiments will be rank-ordered lists, I will measure their similarity using Spearman correlation. As with the first study, it will be important to perform an error analysis to test the model's ability to generalize across topic and platform.

5 Boundary Conditions and Limitations

Misinformation detection is far from a trivial problem, and fully automating the process imposes additional limitations on the methodology. This section will describe these limitations in four major areas: topic generalizability, the types of misinformation that we can expect to detect, and the type of data that can be processed by the pipeline.

5.1 Topic Generalizability

Throughout this proposal, I mention the lack of topic generalizability as a key weakness in the existing work that I cite. This is a common problem because building large, multitopic datasets is resource-intensive and beyond the scope of a typical academic study. As a result, most studies focus on a single topic area (e.g., the CheckThat! competitions focus solely on COVID-19 claims).

The work proposed here attempts to address these issues by expanding to three topic areas, but it is possible that even this is not enough to argue that the methodology is generalizable to any topic. Future work should address this issue, validating the detection pipeline's applicability in a maximally diverse set of topic areas.

5.2 Types of Misinformation

A major limitation of this methodology is the types of misinformation that can be detected. The major limiting factor is the requirement to automate the pipeline in a way that does not rely on human fact-checkers. Without the ability to directly measure the truth value of a message, this methodology will instead look for indirect signals of truth. On the NLP side, that means finding logical fallacies or detecting opposing claims. Some misinformation will certainly escape this detection strategy; however, this work is not aimed at finding all misinformation. Instead, this pipeline is developed to find misinformation that is relevant to social cybersecurity researchers. For misinformation to be relevant, it will have necessarily been spread throughout social media and will undoubtedly encounter resistance from those who disagree (thus, creating an effect in the information space). In other words, we are not looking for claims that are simply false - we are looking for false claims that are having an effect on the information environment.

Other studies have focused more narrowly on propaganda and conspiracy theory detection. As discussed in section 2.4, propaganda detection is not about identifying the truth value of claims and is focused on finding the signals of known methods for spreading claims that are meant to influence a target population in a way that benefits an actor (a country, a political movement, etc.). The methodology in this thesis will likely identify misinformation that is used in propaganda campaigns, but some propaganda is factual information that is shared with a specific intent, which is not something this pipeline is designed to detect. Similarly, this

methodology will detect conspiracy theories that employ misinformation, but in the off chance that a conspiracy theory contains only factual information, we would not expect to detect it.

5.3 Types of Social Media Data

A major boundary condition that I am imposing on this methodology is limiting my studies to English-language data, and there are several reasons this is justified. First, English is my first (and only fluent) language, so focusing on English data will avoid the need for translation or partnering with language experts. Second, the current state of NLP research is focused on English. There are many different multi-lingual models already, and there is no doubt that multi-lingual support is part of the maturation of many existing projects. Fortunately, as things continue to develop towards multi-lingual support in the NLP community, my methodology will be able to adapt rapidly. The work proposed here is essentially a system of existing modeling frameworks that are pre-trained and fine-tuned, but I am not proposing to reinvent any large-scale NLP models. As a result, adding future support for other languages (or general multi-lingual support) will be as easy as substituting the English-only models I used for the desired variants.

Beyond language, I am also limiting this methodology to process text-based social media only. Some of the most popular social media services have moved beyond text to be focused on video and images (notably, TikTok and Instagram), so it is important to note that this methodology is not designed to work on these platforms as proposed. Because this thesis relies heavily on NLP, it will function best on data where text is at the forefront and easy to access. This is not to say that this work will not be applicable to video/image-based social media at some point, but the goal of this thesis is to prove the concept in the ideal case (text data). If it proves to be effective there, future work should look at including support for these more complex cases. One fairly easy extension would involve text extraction from videos and images, while more advanced studies could look to multi-modal approaches that leverage features of the images/videos in addition to the written/spoken text.

A final limitation is driven by the availability of data. The unfortunate trend for social media platforms over the past several years is to close off access to their data. The X API is still available for researchers, but the cost is prohibitively high. Reddit's research API appears to exist, but they have not responded to my requests for access (and from what I can tell, they have not responded to anyone's similar requests). This is not a new problem for Facebook, which has been largely locked behind Meta's expensive API for years. Data from these platforms is still accessible using webscraping techniques, but developing these techniques is a slow and difficult process that ultimately yields a fraction of the data that used to be accessible through the platforms' research APIs.

Due to these significant platform limitations, this work will focus primarily on datasets that already exist on these platforms and data from Telegram. Fortunately, the CASOS lab maintains large datasets containing Twitter and Reddit data from previous work on topics including Elections, COVID-19, and several international events/conflicts. Public Telegram channels are still accessible without limitations from the company's API. In the past, this was not especially helpful to researchers focusing on English-language social media, but the platform is growing in popularity globally, including in English-speaking countries. For research purposes, access to huge amounts of Telegram data will make the platform more useful than access to very small amounts of data from other platforms. Moreover, Telegram is a (mostly) uncensored platform and has been blamed for the spread of misinformation, so a focus on this platform is appropriate for this work.

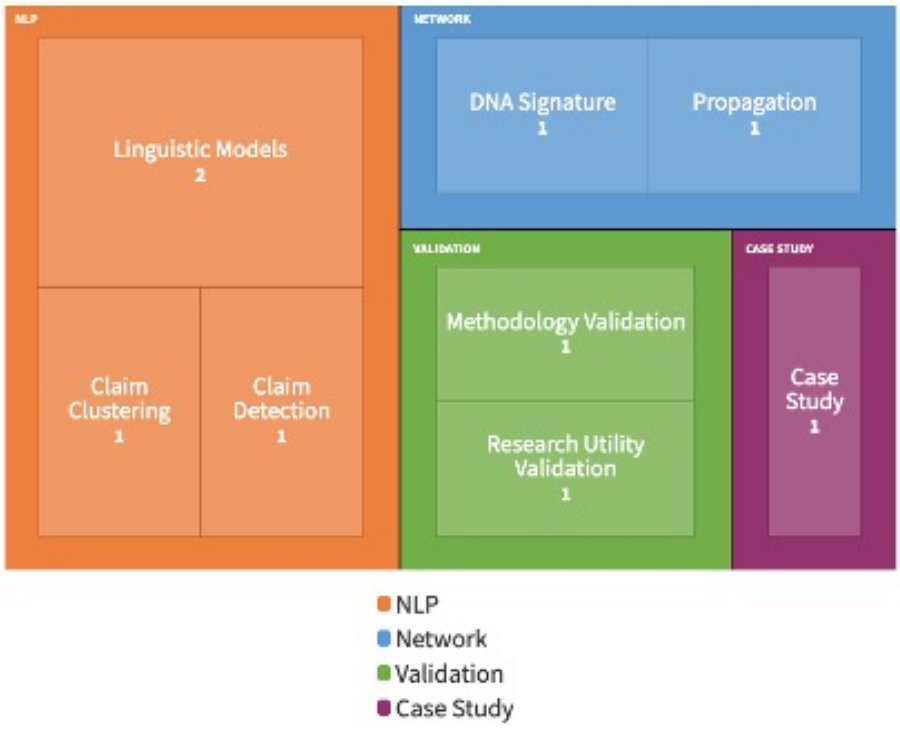

Figure 5: A visual depiction of the contributions from this thesis through NLP studies (orange), network science studies (blue), validation (green), and a case study (purple).

6 Contributions

The main contribution of this work will be a pipeline that can be used to detect misinformation on in large datasets. The goal is for this to be accessible to the average researcher without the need for advanced computation capabilities and rooms full of GPUs. The hope is that this work will be implemented as an ORA report, much like the BEND and stance detection pipelines that currently exists in the software.

In the process of creating the detection pipeline, this thesis will also make contributions to many sub-fields through the defined study structure. Figure 5 shows how these contributions will be divided between NLP, network science, and applied research.

As highlighted in section 3, all of the NLP studies proposed in this thesis are based on existing (and successful) work. The goal is not to reinvent NLP methods, but instead make substantive modifications to the existing work that makes these techniques perform well in a different context. This is more than simply applying existing NLP methods to another domain. Instead, these studies will use processes like pre-training and fine-tuning to produce new models. Both the models themselves and the processes used to make them will be contributions to the NLP field.

Similarly, this thesis does not endeavor to create entirely new network science methodologies but rather to use the existing body of research in new ways. For example, I am not setting out to find a new algorithm for network propagation. Instead, I seek to develop a method for using the existing propagation algorithms in a way that best detects the spread of misinformation. In a similar way, I will not be inventing new dynamic network statistics but will instead search for features that can be used to signal the presence of misinformation in a community.

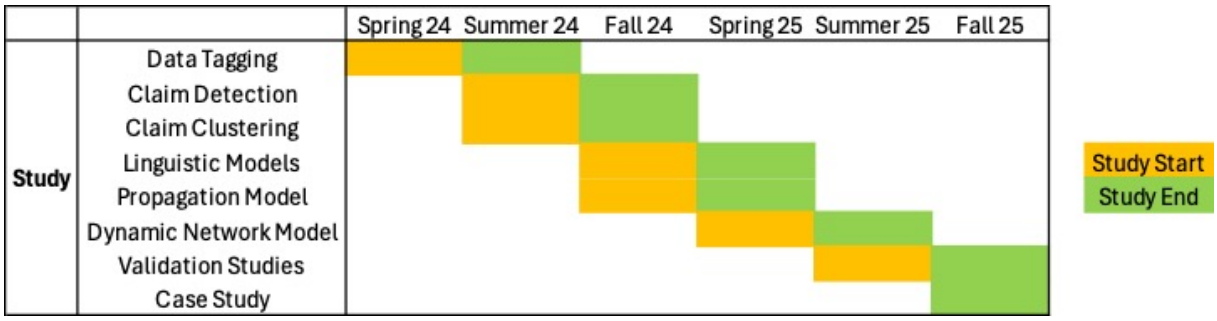

Figure 6: A proposed timeline for completing the work in this document. Yellow rectangles show when work will begin, and green rectangles show when it will be completed.

7 Timeline

Figure 6 shows the planned timeline for this thesis. The first priority is data tagging, as every other part of the thesis relies on it. I plan to finish this work in the summer of 2024 before moving on to the NLP studies. I plan to conclude these in the spring of 2025 before shifting focus to the network studies, which I plan to complete in the summer of 2025. The work will conclude with the validation studies, as well as a final case study that applies the detection pipeline to a new, unlabeled dataset. The subject area of this case study will be selected to be timely and relevant, so I will not propose a specific topic at this point in the research; however, I will select a topic that is prevalent across all four social media services. With this timeline, I will be prepared to defend the thesis at the end of the fall semester in 2025.

None of this work has been completed at the time of this writing. While this is a relatively short timeline, the direction of each study is well-defined (see section 3.1). Furthermore, I have completed large portions of the literature reviews for all of the studies, especially those in the NLP pipeline. As a result, I believe this proposed timeline is ambitious but feasible.

8 References

[1] AllSides. Allsides media bias ratings. `https://www.allsides.com/media-bias`. Accessed: April 1, 2024.

[2] Muhammad Bilal and Abdulwahab Ali Almazroi. Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews. *Electronic Commerce Research*, 23(4):2737–2757, 2023.

[3] Janice Blane. Social-Cyber Maneuvers for Analyzing Online Influence Operations. 5 2023.

[4] Janice T Blane, Daniele Bellutta, and Kathleen M Carley. Social-cyber maneuvers during the covid-19 vaccine initial rollout: content analysis of tweets. *Journal of Medical Internet Research*, 24(3):e34040, 2022.

[5] Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. Finding contradictions in text. In *Proceedings of acl-08: Hlt*, pages 1039–1047, 2008.

[6] Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*, 2022.

[7] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeno, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. Checkthat! at clef 2019: Automatic identification and verification of claims. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*, pages 309–315. Springer, 2019.

[8] Sujatha Das Gollapalli, Mingzhe Du, and See-Kiong Ng. Identifying checkworthy cure claims on twitter. In *Proceedings of the ACM Web Conference 2023*, pages 4015–4019, 2023.

[9] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[10] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412, 2004.

[11] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[12] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[13] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*, 2022.

[14] Vivek Khetan, Somin Wadhwa, Byron C Wallace, and Silvio Amir. Semeval-2023 task 8: Causal medical claim identification and related pio frame extraction from social media posts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2266–2274, 2023.

[15] Ian Kloo and Kathleen M Carley. Social cybersecurity analysis of the telegram information environment during the 2022 invasion of ukraine. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 23–32. Springer, 2023.

[16] Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369, 2017.

[17] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*, 2020.

[18] Lynnette Hui Xian Ng and Kathleen M Carley. Deflating the chinese balloon: types of twitter bots in us-china balloon incident. *EPJ Data Science*, 12(1):63, 2023.

[19] Samantha C Phillips, Joshua Uyheng, Charity S Jacobs, and Kathleen M Carley. Chirping diplomacy: Analyzing chinese state social-cyber maneuvers on twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 95–104. Springer, 2023.

[20] Tiago M Rocha, Ricardo Marau, Tiago Pinto, Rui L Aguiar, and Augusto Matos. Reinforcement learning-based proactive resource management for fog computing. *IEEE Transactions on Network and Service Management*, 17(2):1025–1039, 2020.

[21] Panayiotis Smeros, Carlos Castillo, and Karl Aberer. Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1692–1702, 2021.

[22] Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418, 2023.

[23] Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. Leveraging social discourse to measure check-worthiness of claims for fact-checking. *arXiv preprint arXiv:2309.09274*, 2023.

[24] Twitter. Twitter rules and policies: Manipulated media. `https://help.twitter.com/en/rules-and-policies/manipulated-media`.

[25] Raquel Urena, Gang Kou, Yucheng Dong, Francisco Chiclana, and Enrique Herrera-Viedma. A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. *Information Sciences*, 478:461–475, 2019.

[26] Sander Van Der Linden. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature medicine*, 28(3):460–467, 2022.

[27] Prashanth Vijayaraghavan and Soroush Vosoughi. Tweetspin: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, 2022.

[28] Ferry Wahyu Wibowo, Akhmad Dahlan, et al. Detection of fake news and hoaxes on information from web scraping using classifier methods. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 178–183. IEEE, 2021.

[29] World Economic Forum. The Global Risks Report 2024. `https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf`, 2024.

[30] Fei Xiong, Yun Liu, and Junjun Cheng. Modeling and predicting opinion formation with trust propagation in online social networks. *Communications in Nonlinear Science and Numerical Simulation*, 44:513–524, 2017.