A Hybrid Approach to Polarization Detection: Replacing Traditional NLP tasks with Large Language Models

Alexandros Modestou

Supervisor:

Dr. Marios D. Dikaiakos

Doctoral Student Supervisor:

Demetris Paschalides

Thesis submitted for the award of a Bachelor's degree in Computer Science at the University of Cyprus

Acknowledgments

I would like to firstly thank Professor Marios Dikaiakos for his continued support and feedback regarding the work put in this thesis. It was a very interesting experience and I have learned a lot.

I would also like to thank the Doctoral Student Demetris Paschalides for his support and availability for help, as well as the recommendations put out throughout the thesis.

Abstract

Polarization in media has been escalating in terms of political science and sociological topics. Especially given the surplus of online news and social media, divergent narratives and fragmented worldviews have deepened ideological divides, given rise with the surge of misinformation and distrust. As such, understanding how entities, such as individuals, organizations or institutions, are framed in relation to these topics is essential for capturing the dynamics of this Polarization in media. The POLAR pipeline was introduced as a computational framework to automate the detection and modelling of such polarization by analysing sentiment attitudes between entities and topics within text corpus. However, the POLAR framework is heavily reliant on traditional NLP tools in its initial steps, such as Named Entity Recognition (NER), dependency parsing, and lexicon-based sentiment analysis, all of which suffer from various limitations such as low contextual accuracy and unsupervised error-prone stages. This thesis proposes an augmentation of the POLAR pipeline by integrating state-of-the-art Large Language Models (LLMs) such as GPT-3.5, DeepSeek, and fine-tuned Mistral to improve the extraction of entities, topics, and sentiment attitudes. The enhanced system replaces several heuristic-based components with prompt-based and instruction-following LLM inference. Using a structured dataset of articles, the project introduces a refined methodology for topic identification, entity-pair sentiment classification, and justification extraction for said pairs and attitudes. These improvements feed into the existing POLAR modules such as Signed Attitude Graph (SAG) construction, fellowship detection, and dipole generation, offering more coherent and semantically robust outputs than the NLP steps. As such, to evaluate this integration, the system is benchmarked comparatively to POLAR outputs using metrics such as sentiment classification accuracy, ROUGE-L for justification overlap, and general graph structure. These results indicate that LLMs significantly reduce common error types, particularly dictionary mismatches and dependency parsing issues, while enhancing interpretability and topic coherence. This work demonstrates that replacing static, unsupervised NLP pipelines with prompt-engineered LLMs, not only improves the accuracy of attitude extraction but also enables more nuanced analyses of media discourse, focusing on more sociopolitical topics.

Contents

Chapter 1

Introduction

1.1 Background

Media polarization has increased in the past half-decade, in recent years, the media is plagued with disinformation, populism, fragmentation, distrust and consumer exhaustion. The media is a critical part of the information environment, largely responsible for framing political and social issues and informing the public about key events, often shaping their understanding of key issues in the process. **[2]** Polarization can be defined as a divergence of opinions to opposing ideological extremes, which can be discussed as both a state of being and a process over time. It may also be understood as a behavior, describing how members of a group converge around a specific action such as watching a particular news channel **[3]**.

For the case of this Thesis we want to Highlight 2 types of Polarization, those being Ideological Polarization and Affective Polarization. Ideological polarization refers to the extent to which the groups have divergent beliefs on ideological issues (abortion or affirmative action, etc.) or beliefs that are consistently polarized in a range of issues. Affective polarization refers to the simultaneous positive feelings towards one's own political party or group and negative feelings towards opposing parties or groups, this is most commonly used in political debates.

Polarization has been a prevalent issue in media throughout the world, with many polarizing topics in recent years as the recent 2024 American Elections followed by the later inauguration of Donald Trump as President, with many extremists on both sides feeling strongly opposed to each other, most reflective by Affective Polarization, with each party choosing to support themselves rather than disagree on Policies. Other recent cases such as the 2024 France Olympics introducing multiple layers of discourse, one such case revolving around the Algerian Professional Boxer and Gold Medalist in the women's boxing event Imane Khelif. All sorts of discourse arose regarding her inclusion as well as accusations about her gender worldwide. Another topic of mention is the Brexit vote of 2016, which we will later analyze, which is a historic textbook example of Polarization, with groups of individuals choosing to "Leave" or "Remain", polarizing not just the UK but many countries worldwide in the topic.

In order to Navigate the complex topic that is Polarization, computational approaches are needed to attempt to measure and properly show polarized relationships between all the groups and entities involved, as well as their views surrounding polarizing topics. One such framework, POLAR, was developed in 2021 by Demetris Paschalides **[1]**, for the Modelling of Polarization and Identification of Polarizing Topics. This framework does so by utilizing traditional NLP tools to detect polarization through text from News Articles. However these tools struggle when faced with subtle semantics as well as ambiguity and complexity in text, challenges that modern Large Language Models are considerably better equipped to handle.

1.2 Problem Statement

As mentioned prior, the POLAR framework pipeline offers a structured, rule-based hierarchical method for extracting attitudes and relationships from a corpus of articles, it does so with NLP tools such as Named Entity Recognition (NER) and utilizing sentiment-lexicons (MPQA Sentiment Lexicon **[4]**) for detecting and assigning its attitude values, however this is a static and non-scalable method, which is sensitive to nuance.

As such, with the recent rise of LLMs, capable of understanding and reasoning over long texts, able to analyze sentences contextually given the article's contents, there is an opportunity to

implement its textual parsing and analyzing in order to utilize it, gathering polarizing relationships from text without having to rely on NLP tools.

However, LLMs themselves propose their own sets of challenges, such as having an inherent high computational cost, most often relying on calling external APIs and exhibit significant variability in their outputs.

1.3 Motivation and Mission Statement

The motivation for this project stems from the desire to move beyond rigid NLP pipelines and leverage the full expressive and inferential capabilities of LLMs. By combining the interpretability and structure of POLAR with the power of GPT-3.5 and Mistral, we aim to:

- Detect polarized relationships in a more context-aware and semantically flexible manner.
- Produce explanations (justifications) that are readable, traceable, and informative.
- Enable offline and open-weight inference using a fine-tuned Mistral model.

This thesis proposes a hybrid framework that integrates LLMs into POLAR's architecture and demonstrates the practical advantages of such an approach across multiple evaluation dimensions.

1.4 Objectives

The primary goal of this thesis is to evaluate whether large language models can serve as an effective replacement for POLAR's traditional NLP components in detecting polarized relationships within Articles.

Specific objectives:

- Identify and address the limitations of the traditional POLAR NLP pipeline.
- Integrate GPT-3.5 into the POLAR pipeline for entity/topic extraction and attitude detection.

- Fine-tune the open-weight Mistral-7B model using outputs from GPT to replicate its performance locally.
- Evaluate the effectiveness of both models compared to the original POLAR framework.
- Analyze the trade-offs in terms of model accuracy, speed, interpretability, and operational constraints.

Through this work, we aim to demonstrate that LLMs can significantly improve the interpretability, flexibility, and accuracy of polarization detection in contemporary media.

Chapter 2

Background & Related Work

2.1 Polarization in Media

Media polarization and broader social polarization are distinguished by their focus, measurement, and underlying processes **[5]**. Studies on media polarization concentrate on platform-specific dynamics **[5]**. They define polarization through computable features such as algorithmic curation, selective exposure, semantic divergence, and misinformation **[6]**.

However, research on social polarization emphasizes inter-group conflict, affective responses, and collective narratives **[7]**. Media studies tend to break polarization into interactional, positional, and affective components, while social polarization research centers on broader attitudinal and narrative shifts **[5]**.

2.2 Traditional Approaches to Polarization Analysis

Previous work on media polarization has relied heavily on rule-based and statistical NLP methods. Tools like Named Entity Recognition (NER) utilizing SpaCy **[8]**, noun phrase chunking, dependency parsing, and lexicon-based sentiment analysis have been used to extract entities and assess their stances toward specific topics or one another.

While these techniques offer structured pipelines, they often fail to capture nuance and contextual sentiment, especially in complex or ambiguous language with nuanced sentiments.

2.3 The POLAR Framework

POLAR was developed to automate the extraction of polarized relationships in large-scale corpora of news articles, it is an unsupervised, large scale framework for modeling and identifying polarizing topics in any domain, without prior domain-specific knowledge, it comprises a processing pipeline that analyzes a corpus of an arbitrary number of news articles to construct a hierarchical model graph that models polarization and identify polarizing topics discussed in the corpus. **[1]**

It performs a sequence of NLP tasks to:

- Detect and extract entity-entity and entity-topic relations.
- Assess attitudes (Positive, Neutral, Negative) between these pairs.

However, POLAR's reliance on deterministic, unsupervised components and linguistic assumptions limits its ability to combat implicit sentiment, and complex semantic variation.

2.3.1 The POLAR Pipeline

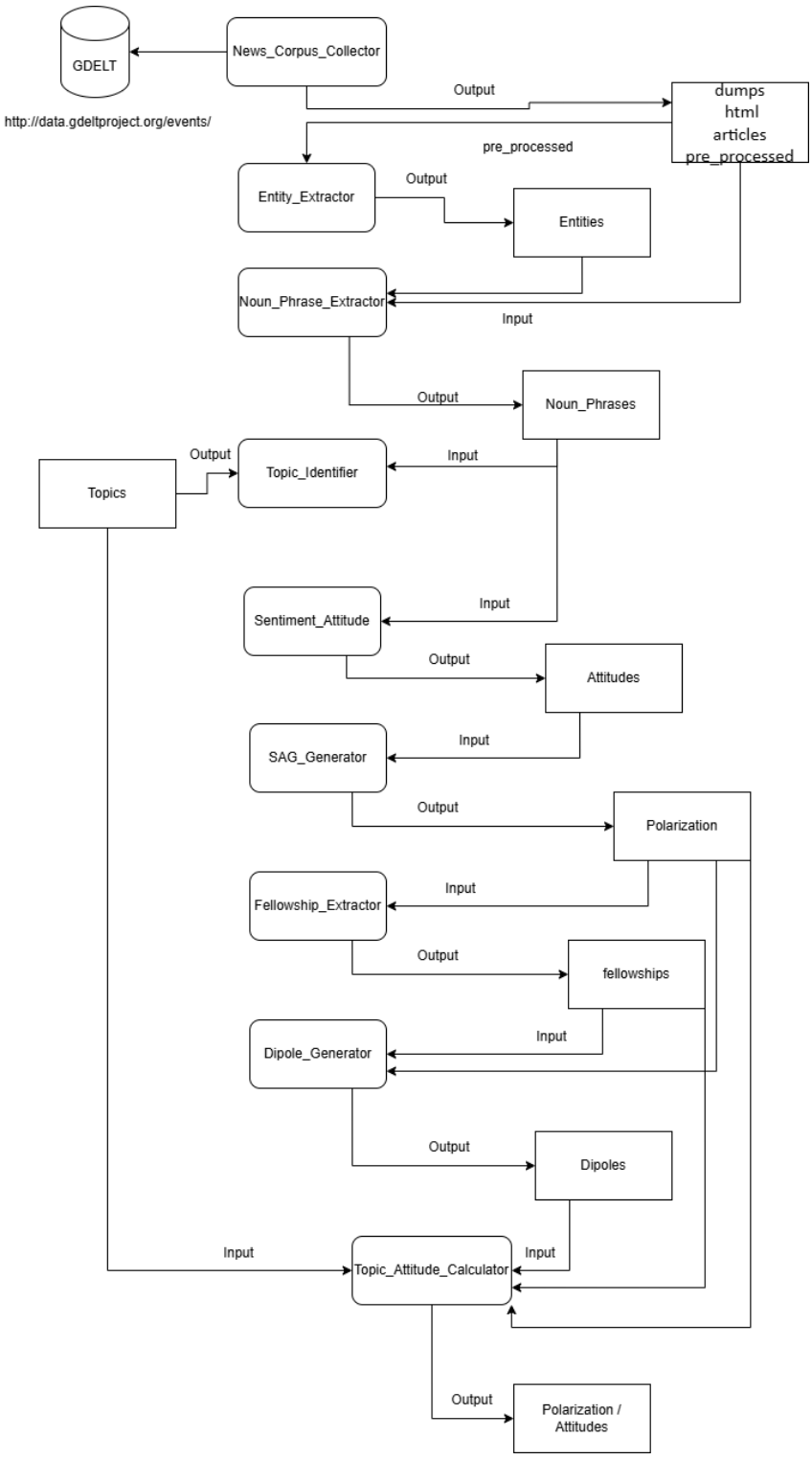

Figure 1 POLAR Pipeline

The POLAR Pipeline, as shown in Figure 1, intends to gather Articles from a News Corpus and, through a series of hierarchical and unsupervised steps, create a Signed Attitude Graph (SAG) comprised of polarized communities as dipoles, and accurate identification of polarizing topics across news media.

The Steps include the **News Corpus Collector**, which, given a set of keywords, the actor countries of origin, the max number of Articles per date and a time period, scrapes the GDELT Database for News Articles that fit the restrictions placed and processes them as Json files locally.

Following the Article extraction is the **Entity Extractor** which processes the Articles, and using NLP tasks like SpaCy locates through the text what it detects as Entities, as well as linking them to a dictionary, in POLAR's case being dbpedia **[15]** with links to detected entities.

The **Noun Phrase Extractor** similarly, using NLP tasks identifies Noun Phrases, however it does not link them onto dbpedia and is much more generous in detecting the Noun Phrases, which are then processed and considered as topics.

The **Topic Identifier** extracts the key phrases from the articles and embeds them with Sentence Transformers, and clusters the similar phrases as topics, scoring and filtering topics based on ther context.

The **Sentiment Attitude** is responsible for analyzing and assigning attitudes between entity pairs and between entities and topics (noun phrases), using a pre-trained transformer to classify the relationships as Positive, Neutral, Negative.

The **SAG Generator** (Signed Attitude Graph) constructs a graph representation of inter-entity relationships based on the previous sentiment analysis results, by loading the sentiment attitude between entity pairs from the previous steps and building a SAG where nodes are the Entities and Edges are the Sentiment between them.

The **Fellowship Extractor** identifies clusters of entities that show strong positive interrelations in the SAG and detects communities or groups of closely aligned actors based on the polarity of the positive edges.

The **Dipole Generator** identifies pairs of fellowships that are in strong opposition to each other, this is done so by utilizing the connections made in the Sentiment Attitude layer, the Dipoles range from having Polarizing connections to Agreeing connections to Neutral.

The **Topic Attitude Calculator** is the final step of the pipeline, and shows how the different Fellowships interact with others towards different topics, showcasing the Dipoles in a SAG format. The Topics in question are labeled as Positive, Neutral or Negative from either Fellowship in the Dipole.

2.3.2 POLAR Limitations

2.3.2.1 Entity/Noun Phrase Extraction:

These initial steps are used in order to extract Named Entities from the set of Articles from the Corpus, as mentioned previously in the Description of each step. It utilizes Named Entity Recognition (NER) using SpaCy's built-in NER component **[8]** in order to identify the Entities from the text this way, as well as later normalize them by linking them to a Dictionary that matches the entity. The Noun Phrases are similarly parsed this way using SpaCy's Dependency Parser, detecting what is then represented as topics.

These steps are crucial to the POLAR Pipeline, as they provide the foundation for the following steps, as POLAR is an unsupervised hierarchical framework, it assumes all of the layers are executed properly and depends on them this way, meaning any mistakes or loss of nuance in a step will affect the capabilities of the later steps processing them, for example detecting false relationships with false sentiment will affect the SAG in a negative way as well as the later dipoles.

The initial steps mentioned are error prone, due to the natural limitations of these NLP tasks, the most common error types being Dictionary Errors and Dependency Parsing Errors. These errors are often found in nuanced text with complex sentiment, and it is not rare to detect POLAR mistakenly attributing wrong Dictionary Links, or failing to grasp the Sentiment through pure textual Syntax, with some mistakes being common per article analysis.

Dictionary Errors:

This type of error occurs when an NLP component like POLAR's Normalization step, or POLAR's usage of the MPQA library to apply Sentiment, relies on a fixed lexicon or a static way of assigning components and fails to interpret a word or phrase.

Dependency Parsing Errors:

this refers to a mistake in the syntactical analysis of a sentence, where the parser incorrectly identifies the relationship between words, resulting in incomplete noun phrases and mistaken attitudes.

Example:

entity1: "Howard" -> reference: "http://dbpedia.org/resource/John_Howard"
entity2: "Schneider" -> reference: "http://dbpedia.org/resource/Schneider_Electric"
Sentence: "Participants hold a British Union flag and an EU flag during a pro-EU referendum event at Parliament Square in London Thomson Reuters By David Lawder and Howard Schneider Advertisement WASHINGTON (Reuters)"

***Problem**: Assigns The Entity Pair where Entity1: "Howard". and Entity2: "Schneider", when in reality, "Howard Schneider" is the full name of the person, mistakenly identifying a pair that should not exist.*

Dependency Parsing Error: Incorrectly splits the Entity "Howard Schneider" into 2 subparts "Howard" and "Schneider" and assigns different dbpedia links to them ("John_Howard" and "Schneider_Electric").
Dictionary Error: Does not correctly link the Entity "Howard Schneider" to a pre-existing dbpedia link, as it does not exist.

2.3.2.2 Sentiment Attitude Extraction

This step estimates the sentiment between pairs of entities or noun phrases based on the syntactic paths that connect them in the sentence and then assigns a polarity score to them, being Positive, Neutral or Negative.

It utilizes techniques such as Dependency Parsing via SpaCy, and Lexicon-based sentiment using the Subjectivity Lexicon provided by Multi Perspective Question Answering (MPQA), which is a lexicon manually describing the Sentiment of words and Noun Phrases in the format of "Word"="Sentiment" per-line. **[4]**

This step is also highly error prone to Dependency Parsing errors, as the entire process hinges on correct dependency parsers, and is not very effective against complex nuance, sarcasm etc.

Example of Dependency Parsing Error:

entity1: "U.S. monetary policy" -> reference: "http://dbpedia.org/resource/Monetary_policy_of_the_United_States"
entity2: "British" -> reference: "http://dbpedia.org/resource/United_Kingdom"
Sentence: " A vote by Britons to leave the European Union on Thursday may not drag the United States into recession, but its effects on U.S. monetary policy, trade and corporate profits are causing concern in Washington D.C. and boardrooms alike."
Attitude: "Positive"

***Problem**: Attitude between the "U.S. monetary policy" and the "British" is labeled Positive however the Sentence describes how it is causing concern about its effects possibly going into a recession, this should be a Negative Pair given the Sentence.*

2.4 The Rise of Large Language Models

Recent advancements in large language models (LLMs) have led to remarkable gains in natural language understanding **[17]** . Models like GPT-3.5 and Mistral-7B demonstrate state-of-the-art performance in tasks involving sentiment detection, contextual reasoning, summarization, and named entity recognition **[17]** . Their capacity to generalize from context makes them compelling candidates to replace rule-based systems, such as the aforementioned, error prone modules in POLAR.

Chapter 3

Large Language Models and Mistral

3.1 Introduction

This chapter provides an overview of the large language models used in this thesis: GPT-3.5-turbo and Mistral-7B, as well as the attempts made at utilizing Llama and Deepseek. It outlines their architectural foundations, capabilities, and their ability to replicate POLAR's NLP tasks (Entity Extraction, Noun Phrase Extraction, Noun Phrase Attitudes, Syntactical Sentiment Attitude).

While attempting to run the LLM Integrated pipeline, we tried these options by modifying the payload, variables and window size as well as any API needed. The LLMs tested were all used during the LLM Communicator part of our Methodology Pipeline, see Chapter 3 Figure 2.

3.2 GPT-3.5-turbo

GPT-3.5-turbo is a powerful closed-source model developed by OpenAI, optimized for fast and cost-effective inference via API. It supports instruction-following tasks and demonstrates strong performance on reasoning, Entity recognition, and Sentiment tasks. **[12]**

For our calculations, GPT was used as the benchmark for the other LLMs, as it followed the prompt well and gave concise responses that accurately predicted entity pairs and topical pairs if it detected any in the articles. The speed of the replies of GPT were also the highest in terms of the LLMs, mostly due to the fact it was called via an API, which also allowed for some Parallel calling, the Numbers being as follows:

Dataset of 563 Article Parts:

GPT:

-GPT_Communication Elapsed time: 10280.43 seconds == 171.34 minutes == 2.856 hours

Polar:

-entity_extractor Elapsed time: 640.87 seconds

-noun_phrase_extractor Elapsed time: 4545.63 seconds

-topic_identifier Elapsed time: 1226.28 seconds

-sentiment_attitude_pipeline Elapsed time: 239.86 seconds

-load_sentiment_attitudes Elapsed time: 2.89 seconds

-> 6653.53 seconds == 110.89 seconds == 1.848 hours

Polar was faster for the dataset when compared to a sequential model of requesting the API of GPT, with each GPT request taking about 18 seconds with around 5 seconds of standard deviation, since the request is directly tied to the length of the article and the Entity/Topical Pairs detected, from our experience, the code is partially parallelizable and it overall decreases the overall time needed, however runs the risk of blocking the API with too many calls depending on the requests per second.

3.3 Deepseek

The other LLM attempted which yielded promising results was the currently newly implemented Deepseek, specifically Deepseek-r1-8B, which is a recent open-source model developed by Deepseek excelling at a wide range of tasks like instruction following, reasoning and knowledge

retrieval**[13].**

Using the same technique used in GPT with the same format, Deepseek returned very similar results that were properly structured as specified by the prompt, while also being fully localized, however the time per Article was tanked as a result:

-Average time required Per Article Part: 900 - 1800 seconds == 15 – 30 minutes

An attempt was made to Parallelize the code to test whether it would help, however since it was being run locally on the device, the overall time did not improve:

-Average time required Per Article with 4 Threads, where each Thread operates on each own Article: 2700 – 3500 seconds == 45 – 58 minutes

Due to the time required per article, we did not further pursue using Deepseek, even if the short number of Articles parsed were satisfactory.

3.4 Unsuccessful Implementations

Llama:

A small attempt was made to utilize Llama, we tried to install it locally through the meta instructions but were unsuccessful, we then attempted to use Llama through Groq **[16]** by using their API service. The results from Llama were unsatisfactory, with the LLM often not following the prompt instructions well as well as not outputting in Json format most of the time, the Groq API also would also limits its service to about 300 Articles, limiting its usage.

Mistral:

We also tried to Use Mistral, which at first was unsatisfactory, not following the prompt despite the detailed prompts. However, after fine-tuning it with GPT's outputs, it became satisfactory, almost replicating GPT's outputs, Evaluation results shown in later Chapters. The Process of the Fine-Tuning is mentioned in Chapter 4.4, using all the Responses gathered from GPT as its training data with very positive results.

Chapter 4

Methodology

4.1 Overview

This chapter describes the proposed methodology for replacing POLAR's traditional NLP components mentioned, with a unified, LLM-powered approach.

The new system leverages the language understanding capabilities of LLMs to extract Entity-Topic and Entity-Entity relationships, (where Entities would correspond to Actors, Groups, Nations) determine their attitudes, and justify those relationships, all in a single request using structured prompts. In our cases all of the data extracted by the use of LLMs was done so using GPT-3.5 turbo, we additionally fine-tune an open-weight Mistral-7B model to replicate this behavior locally.

This Implementation is readily available for use via the github link, provided an API key or local LLM.

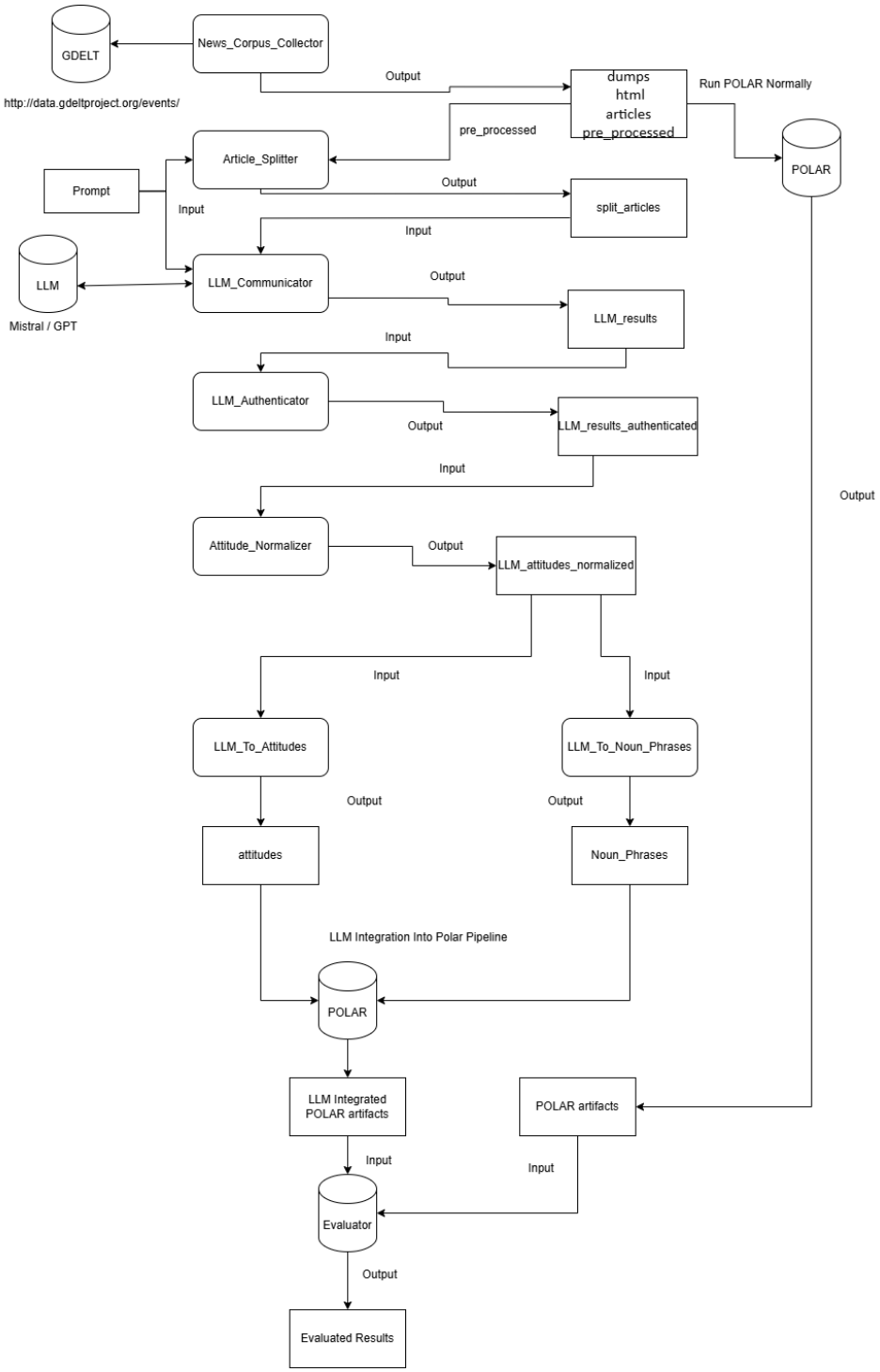

Figure 2 LLM Integrated POLAR Pipeline

4.2 Data Collection and Preparation

The dataset used was collected using POLAR's News_Corpus_Collector module, which scrapes relevant news articles over defined date ranges and topics. The Articles gathered this way are filtered for length, language, actor location, and source credibility through POLAR.

The gathered Articles are then split into paragraphs or segments of manageable length, with the intent to comply with the token limits set for the window sizes of each respective LLM called. This is done using LangChain **[11]** to detect proper locations to break the Article Text, prioritizing Stop words like commas and periods, the Articles are Split into Segments such that each Segment + the Prompt for the LLM does not exceed the Window Size set. These Article segments are then passed to the LLM for a structured parse and analysis.

4.3 Prompt Based Extraction

We construct a structured prompt that instructs the LLM to do the following:

- Identify Entities (politicians, groups, nations etc.) as well as relevant topics from the Article.
- Determine The Entities' Sentiment Attitude towards the other Entities or Topics based on the Article contents.
- Provide textual evidence based on the Article Content to Justify the relation pairs detected above.

It is important that the prompt is not ambiguous or has any room for misunderstanding in its instructions, such that the curated responses are as similar as possible and able to be automatically parsed and processed without human intervention. Due to this, the length of the prompts needed are quite lengthy, as well as containing an example of the structured expected, as it boosts consistency in the responses. It is important to note that the more restrictive the prompt is in its instructions reduces the variability of the LLM, making the outputs more consistent and as such excluding less Responses in the Validation. The lengthier prompts however take more and more space away from actual Article content.

An example Alpaca-style prompt template is used as payload for the LLM:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:

Our detailed Prompt that instructs a specifically formatted response with detailed instructions.

Input:

[The Article Segment to be Parsed]

Response:

[The LLM's response in processable Json format]

The requests are sent sequentially, where each Article Segment has its own unique interaction with the LLM, this has the benefit of having a minimal amount of Threading available, provided the requests do not cause the API to block the requests. However, having all of the requests take their own independent request means that any nuance from past Article Segments gets lost. This is a weakness of the current implementation and can be fixed in the future by detecting when an Article was Split and keeping the connection for that Article, where the LLM would remember the whole conversation, this could also lead to not needing to resend the prompt in these circumstances.

The LLM response is then validated to match our expected Json structures, specifically for it to include the entity_attitudes and topical_attitudes fields, of which they contain the Entity to Entity Relations and Entity to Topic Relations accordingly. The files are validated by removing any non-Json formatted text in the reply, this includes ```json ``` blocks as well as any additional comments from the LLM. Then the Response is checked for the appropriate fields in both entity and topical attitudes, such that they can then be modified and parsed accordingly.

An attempt is then made to normalize the reference fields of all entities and topics among the articles, in order to couple up references to the same entity and topics (President Obama -> Obama, etc.), this is done using semantic similarity, as well as a manual alias table for the topic.

The Normalized Responses are then converted to POLAR's Noun Phrases and Attitudes fields accordingly and the Pipeline proceeds as normal.

4.4 Fine-Tuning Mistral

While for the process of data gathering and testing, GPT-3.5 was effective and efficient, it is closed-source and requires paid API access. For scalability and local use, as well as potential improvements, we fine-tune the open-source Mistral-7B model using the same data generated from GPT-3.5 responses.

Fine-Tuning Process:

Using the Unsloth library **[14]**, we apply Low-Rank Adaptation (LoRA) to fine-tune a quantized version of Mistral-7B (*unsloth/mistral-7b-bnb-4bit*) on formatted prompt-response pairs. Alpaca-style instruction formatting is used, and training is conducted using Hugging Face's SFTTrainer with LoRA and 4-bit quantization for efficiency.

Model parameters:

- LoRA rank: 16
- Max sequence length: 4096
- Training steps: 60
- Optimizer: AdamW 8-bit

The Dataset consisting of the generated replies of the GPT model, as well as multiple similar Alpaca-style instructions generalized towards gathering Entity Relations towards topics. Due to the nature fine-tuning datasets, it is possible to skip the lengthy prompt needed for other LLMs, by prompting a simple minimal Instruction with each given response, by essentially telling the model that this is what the other LLMs output given that prompt, and by doing so for thousands of instructions it manages to replicate the instructions of the lengthy prompts. The result being a local LLM capable of replicating GPT's responses.

It is possible that with a bigger dataset and more specified fine-tuning that a fine-tuned model specifically trained for gathering polarization could outclass the other LLM options.

4.5 Summary

This methodology section has outlined our full pipeline: from data collection and preparation to prompt engineering, API interaction, and local fine-tuning. The next chapter describes the models used in more detail, followed by the details regarding Evaluating the performances.

Chapter 5

Evaluation

5.1 Overview and Objectives

This chapter outlines the methodology used to evaluate the different systems designed to extract and represent polarization, including the fine-tuned Mistral model, the original LLM Integrated POLAR Pipeline, and the standard POLAR system.

The goal of this evaluation is to determine how well each system can identify relevant Entity-Topic / Entity-Entity relationship pairs and predict their associated attitudes and justifications for said Attributes and pairs.

Given the inherently subjective and interpretative nature of polarization, traditional classification metrics are complemented by semantic and qualitative analyses as well as further analyses and measures in Chapter 6.

5.2 Evaluation Pipeline

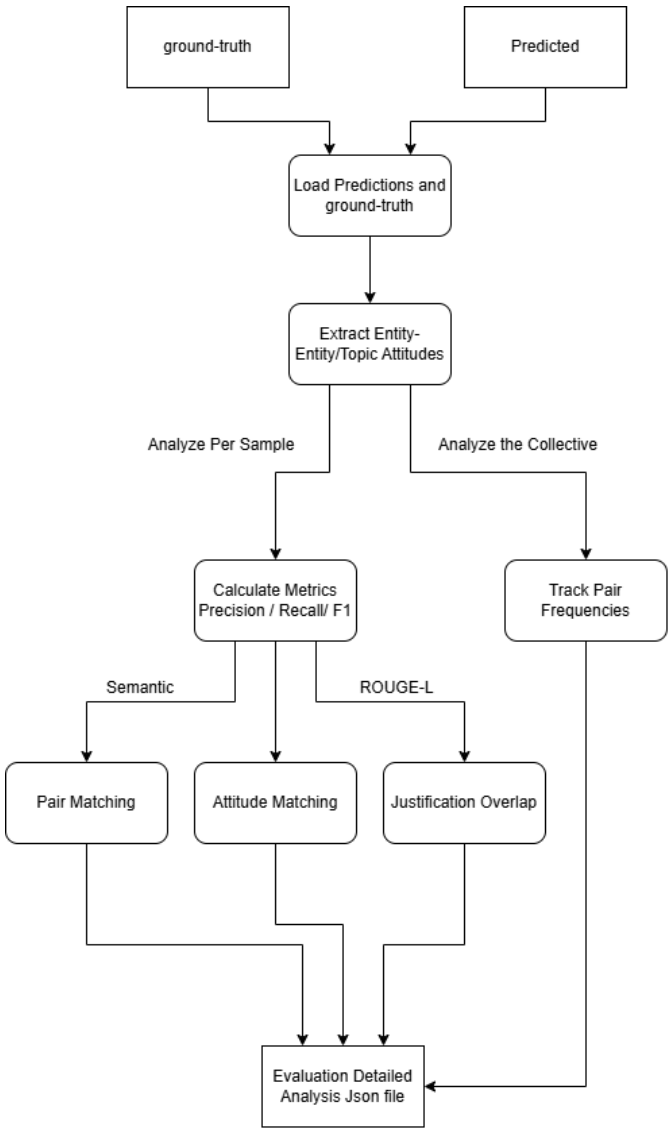

Figure 3 Evaluation Pipeline

To systematically compare systems, we designed a modular evaluation pipeline. Figure 3 illustrates the sequence of steps applied to each pair of ground-truth and predicted Json files:

1. **Load Data**: Import and parse the files of both the ground-truth and predicted set into Json structure and order them such that each entry of the predicted and ground-truth set refer to the same Article.

2. **Pair Matching**: Identify Entity-Topic and Entity-Entity pairs in both sets, and calculate matches using Semantic Similarity with Sentence Transformers **[9]** , comparing each pair from the predicted set with each ground-truth pair, until a match is found or it checks every pair, repeating for each predicted pair. A matched pair being considered when both Source Entities match and both Target Entities/Topics match, using a threshold of 0.7.

   Example matches using Sentence Transformers:

   Matching Topic Pair: (Britain, United Kingdom)
   Matching Topic Pair: (Ukraine, Kyiv)
   Matching Entity Pair: (Tony Blair, British prime minister)
   Matching Entity Pair: (Trump, Donald Trump)
   Matching Entity Pair: (Britain, France)
   Matching Topic Pair: (Germany, Russia))

3. **Attitude Comparison**: For all matched pairs, compare the predicted attitude (Positive, Negative, Neutral) against the ground truth, considering the matches as True Positive and the wrong predictions as False Negatives.

4. **Justification Evaluation**: For all matched pairs, ROUGE-L **[10]** is called to compare textual overlap between predicted and true justifications, considering matched Sentences those whose overlap reaches a threshold of 0.5.
   ROUGE-L is a metric that measures how similar two sequences of text are based on their Longest Common Subsequence (LCS), considering that if two pieces of text share a long ordered subsequence of words, they are semantically, or structurally similar, with a higher ROUGE-L score indicating greater textual alignment. Especially in our case, since the LLMs and POLAR quote the Article directly in the justification field, if they have similar subsequences, they are likely referring to the same sentence from the Article.

5. **Pair Frequency Analysis**: Count the number of unique Entity-Topic and Entity-Entity pairs and count each of their appearances in both the ground-truth and the Predicted Set. It does so by Normalizing the fields:

   Lowercasing the fields
   Removing Punctuation like commas and periods

Singularizing the fields, turning plural nouns into singular

Lemmatizing the fields, converting words to their basic forms

Manually aliasing certain words (trump->donald trump, eu->european union)

This is done so in order to more accurately cluster mentions of certain entities throughout multiple Articles and group up certain similar topics.

This showcases the appearance rate of each pair, as well as the frequency of which it is assigned each attitude by both models.

5.3 Metrics Used

Pair Matching Metrics

To assess how well the Predicted dataset identifies the ground-truth Entity-Topic or Entity-Entity pairs, we compute:

- **True Positives (TP)**: Pairs present in both predicted and true sets and correctly matched.
- **False Positives (FP)**: Pairs present in predictions but not in the ground truth.
- **False Negatives (FN)**: Pairs present in ground truth but missed by predictions.

Using these metrics we can calculate:

- **Precision**: Precision = TP / (TP + FP). Measures the accuracy of the predicted pairs, whenever the prediction detects a pair, how often it is found in the ground-truth.
- **Recall**: Recall = TP / (TP + FN). Measures the model's coverage of all relevant pairs, shows the percentage of missed pairs undetected from the ground-truth.
- **F1 Score**: F1 = 2 * (P * R) / (P + R). Harmonic mean of Precision and Recall.

Attitude Prediction Metrics

For all matched pairs, we evaluate the similarity of their Attitude Sentiment:

- **True Positives (TP)**: Sentiment Attitude of Predicted pair matches ground-truth pair

- **False Negatives (FN)**: Sentiment Attitude of Predicted pair does not match ground-truth pair

Using these metrics we can calculate:

- **Accuracy:** TP / TP + FN. How accurate the Predicted model is at matching the ground-truth model.

Justification Overlap

For all matched pairs, we use **ROUGE-L** to calculate the textual alignment between the two models and see whether it reaches the threshold (0.5) to be considered a match, using these metrics:

- **Matched Justifications (MJ)**: The ROUGE-L score calculated reaches the threshold, the justifications match.
- **Total Justifications (TJ)**: All the Justifications present in the Matched Pairs.

As such we can calculate:

- **Justification Overlap:** MJ / TJ. The percentage of the Total Justifications that were successfully matched between both pairs.

Following this is the results of the Evaluations of POLAR with GPT, as well as the Evaluation of the Fine-Tuned Mistral-7B Model.

Chapter 6

Results And Analysis

6.1 Overview

6.1.1 What is being evaluated

This chapter evaluates the artifacts produced by integrating Large Language Models (LLMs) into the POLAR pipeline. Specifically, it assesses how well GPT-3.5 and a fine-tuned Mistral-7B model replicate or improve upon POLAR's traditional NLP pipeline. The artifacts evaluated include Entity and Topical attitudes, Signed Attitude Graphs (SAGs), Fellowship groupings, Topic coverage, and Attitude Dipoles, on the following systems:

- **POLAR**: The baseline pipeline using conventional NLP tools.
- **GPT-3.5**: A large language model used as a drop-in replacement for several POLAR components.
- **Fine-tuned Mistral-7B**: An open-weight language model trained to replicate GPT-3.5's outputs and evaluated for fidelity, efficiency, and practical deployment advantages.

6.1.2 Metrics and Qualitative Measures Used

Evaluation metrics were computed over the subset of articles for which both the GPT-generated ground truth and the Mistral predictions produced valid Json structured outputs. This excluded any malformed responses from both GPT and Mistral for their respective datasets.

Pair matching and attitude evaluation were performed on a per-article basis, with each sample evaluated independently. As such, the same entity or topic pair may appear multiple times across the dataset and is evaluated separately in each occurrence.

Metrics:

In the context of Pair Matching:

“True Positive” : X number of pairs from the Predicted Dataset were matched with pairs from the Ground Truth Dataset
“False Positive” : X number of pairs were found Only in the Predicted Dataset and were not included in the Ground Truth Dataset
“False Negative” : X number of pairs were found Only in the Ground Truth Dataset and were not included in the Predicted Dataset

In the context of Attitude Matching:

“True Positive” : X number of Attitudes were correctly predicted to the Ground Truth Attitude
“False Negative” : X number of Attitudes were incorrectly predicted to the Ground Truth Attitude

In the context of Justification Overlap:

Overlap: How similar the model’s justification for the Pair/Attitude were, using ROUGE-L **[10]** for the calculation of the similarity with a threshold of 0.7 to be considered matched.

6.2 POLAR Compared to LLMs Evaluation

This section will focus on the evaluation of the GPT Integrated Pipeline when compared to the Standard POLAR pipeline.

6.2.1 Attitudes Comparison

6.2.1.1 Attitudes Evaluation

In this section of the evaluation, we aimed to assess the extent to which the POLAR system could replicate the Attitudes field generated by a large language model (LLM), specifically GPT-3.5. The Attitudes field comprises entity or topic relationships and the corresponding

sentiment classification (Positive, Neutral, or Negative). The goal was to analyze the degree of similarity between the attitudes identified by POLAR and those generated by GPT-3.5 for the same input samples.

To conduct this comparison, we designated GPT-3.5's output as the reference (or ground truth) and applied POLAR to the same dataset, which consisted of approximately 404 article segments.

The Attitudes fields produced by POLAR were then treated as predictions and compared against the GPT-3.5 outputs to evaluate alignment and divergence.

Evaluation - Polar completed in 395.37 seconds. (~6.59 minutes)

"Entity"

Pair Matching

Precision	*0.0787*
Recall	*0.0295*
F1	*0.0429*
True Positives	*20*
False Positives	*234*
False Negatives	*658*

Attitude Prediction

Accuracy	*0.35*
True Positives	*7*
False Negatives	*13*

Justification Overlap

Overlap	*0.1379*
Matched Justifications	*4*

Total Justifications	*29*

From this data, we can see that from the evaluated dataset, POLAR Successfully manages to predict 20 pairs from the 678 pairs in the Ground Truth dataset == roughly 2.95% of the pairs in Ground Truth were accurately predicted.

Of those 20 pairs, 20 attitudes were compared between the Ground Truth and the Predicted Dataset, with POLAR correctly predicting 7 of them == roughly 35% pairs were correctly attributed. Of those 20 pairs, the justification was successfully matched using ROUGE-L similarity roughly 13.79% of the time.

"Topical"

Pair Matching

Precision	*0.0300*
Recall	*0.0528*
F1	*0.0383*
True Positives	*44*
False Positives	*1422*
False Negatives	*789*

Attitude Prediction

Accuracy	*0.1591*
True Positives	*7*
False Negatives	*37*

Justification Overlap

Overlap	*0.0*
Matched Justifications	*0*

Total Justifications *0*

From this data, we can see that from the evaluated dataset, POLAR Successfully manages to predict 44 pairs from the 833 pairs in the Ground Truth dataset == roughly 5.28% of the pairs in Ground Truth were accurately predicted.

Of those 44 pairs, 44 attitudes were compared between the Ground Truth and the Predicted Dataset, with POLAR correctly predicting 7 of them == roughly 15.91% pairs were correctly attributed. The Justification Overlap was overlooked for the Topical case here since POLAR does not provide a Sentence for its Noun_Phrase_Attitudes, which are what the Topical Pairs here are.

6.2.1.2 Attitude Comparison Conclusion

From these results, POLAR and GPT, in terms of a per-article basis, have widely different results each, with POLAR failing to emulate GPT's output.

In terms of Entity Pairs, POLAR matched 20 of its 254 results to GPT, this shows that POLAR is more conservative with what it considers an Entity Pair, when compared to the 678 GPT results. From the Precision we can see that when a POLAR Entity Pair is calculated, it will match GPT 7.87% of the time.

However, the Topical Pairs calculated by POLAR are nearly double GPT's, with 1466 vs 833 respectively, however only managing to match 44 of them (5.28%), given the precision, we see that when a POLAR Topical Pair is calculated, it will match GPT 3.00% of the time.

In terms of Attitude Prediction and Justification Overlap for the matched pairs, POLAR is also dissimilar to GPT, only having 15-35% similar Pair Attitudes with only roughly 13-14% Justification Overlap.

6.2.1.3 Discussion

These results do Not tell us that POLAR is a worse model than GPT, or vice versa. All this means is that fundamentally, for a per-article basis, these models are completely different in what they consider Pairs and calculating Attitudes.

The issue regarding comparing the calculated Attitudes for a line of text, is that there is no inherent "Absolute", or "Correct" amount of Pairs to expect these models to output, the same can be said about the attitude between the Entities/Topics to be calculated merely through text.

6.2.2 Further analysis

From an initial comparison, the models differ drastically when comparing their results per article, however we wanted to see their results Overall as well as compare their results in different parts of the POLAR pipeline instead of just their Sentiment Attitude Assignment.

6.2.2.1 Overall Attitude Summary

These graphs show the Overall number of Positive/Neutral/Negative attitudes that GPT assigned to the pairs, as well as the Seaborn Graph showcasing the amount of those attitudes assigned for each sample, where each Dot corresponds to one Sample.

Training Dataset:

"True_Data"(GPT):

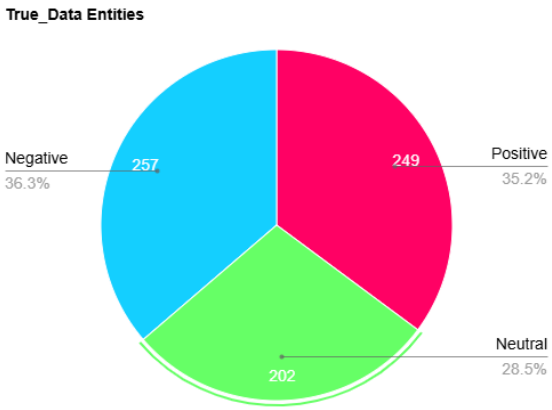

Figure 4 Overall Attitude showcase True Data Entities

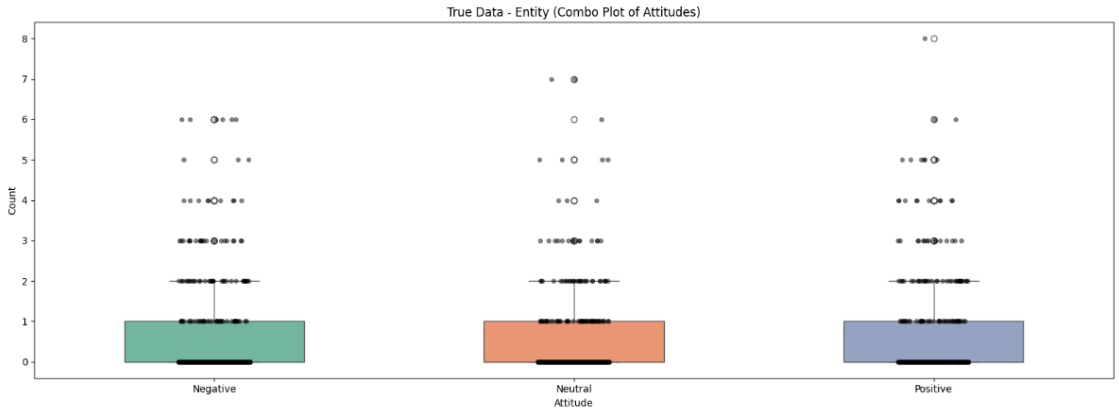

Figure 5 Overall Attitude Seaborn Graph True Data Entities

These Graphs show us that for this dataset comprising of Articles from multiple topics, GPT covers both Positive/Neutral/Negative consistently in terms of Entity Pairs, averaging 0-1 of each per Sample with the highest number of assigns being 8 Positive for a single Sample.

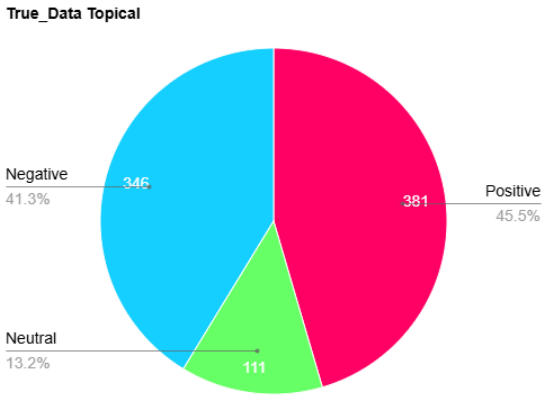

Figure 6 Overall Attitude showcase True Data Topical

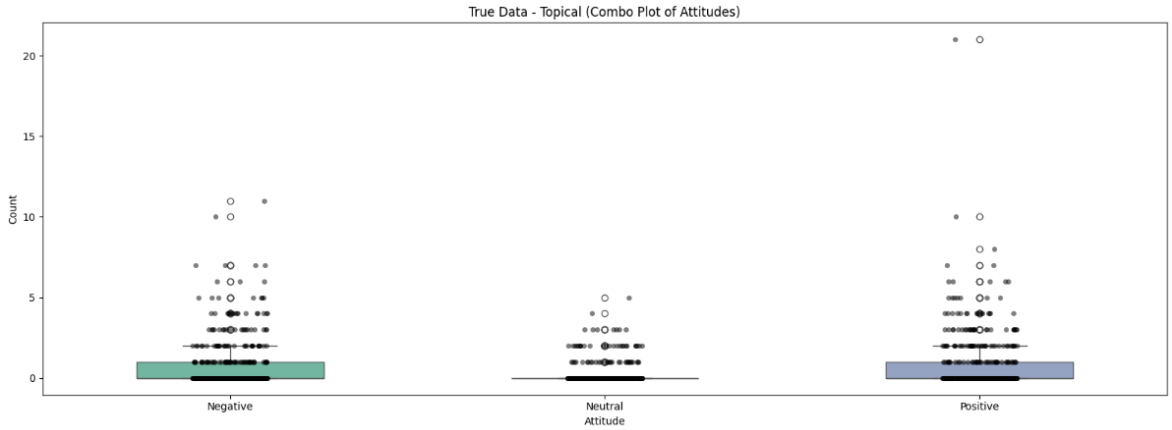

Figure 7 Overall Attitude Seaborn Graph True Data Topical

These Graphs show us that for Topical Pairs, GPT covers both Positive /Negative consistently with a few Neutral assigns, averaging 0-1 of each per Sample except Neutral with the highest number of assigns being 20 Positive for a single Sample.

"Pred_Data" (POLAR):

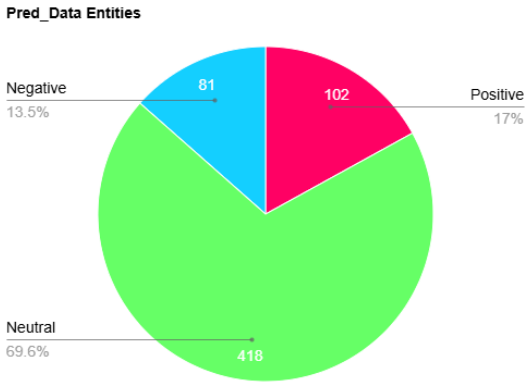

Figure 8 Overall Attitude showcase Predicted Data Entities

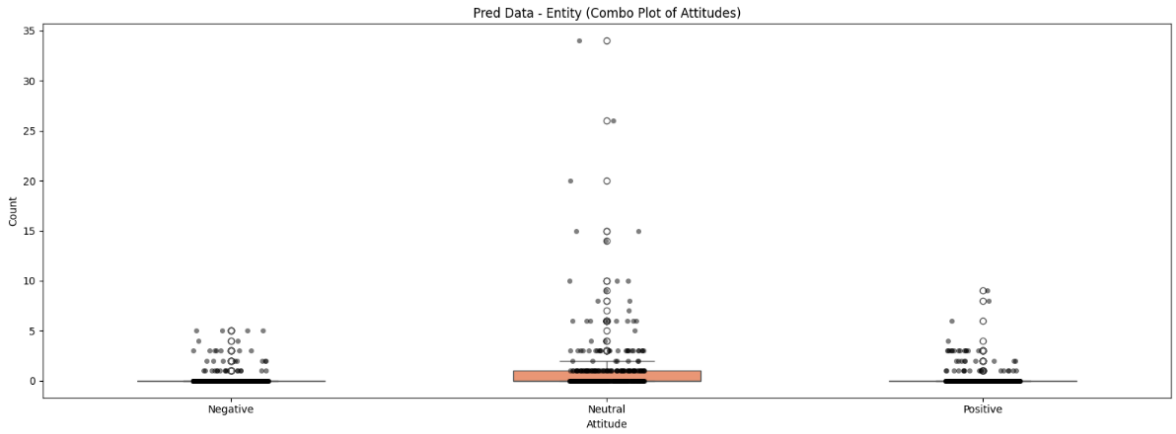

Figure 9 Overall Attitude Seaborn Graph Predicted Data Entities

These Graphs show us that for Entity Pairs, POLAR covers a large majority of Neutral pairings, with some Positives/Negatives, averaging ≈0 Positive/Negative and 0-1 Neutral assigns per Sample, with the highest being 35 Neutral assigns for 1 sample.

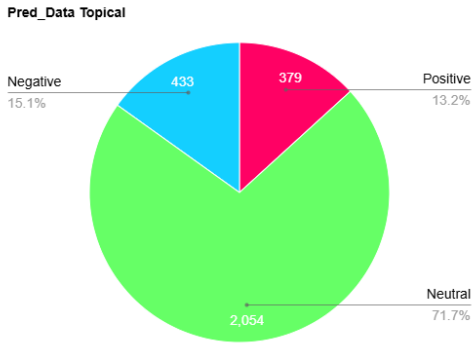

Figure 10 Overall Attitude showcase Predicted Data Topical

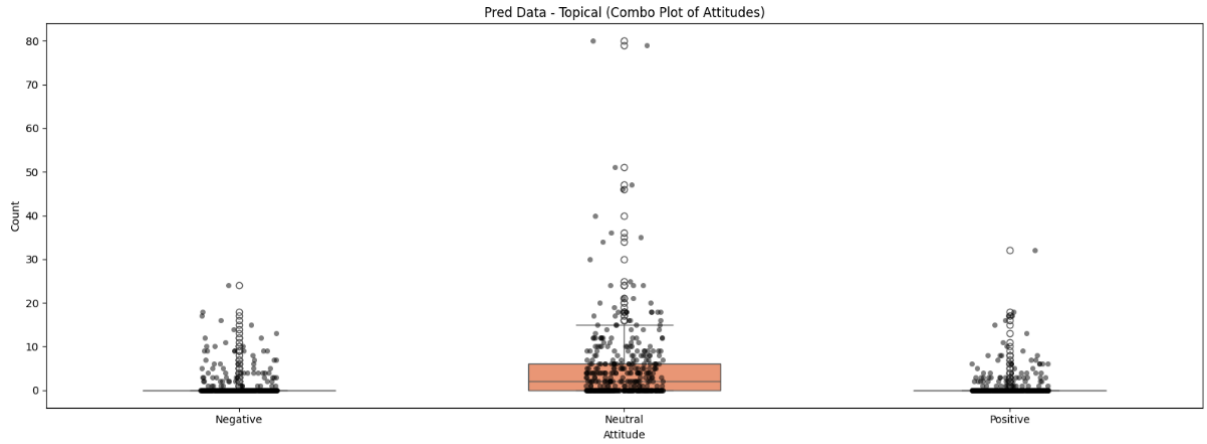

Figure 11 Overall Attitude Seaborn Graph Predicted Data Topical

These Graphs show us that for Topical Pairs, POLAR covers a large majority of Neutral pairings, with some Positives/Negatives, averaging ≈0 Positive/Negative and 0-7 Neutral assigns per Sample, with the highest being 80 Neutral assigns for 1 sample.

Brexit Dataset:

An additional analysis was conducted on a dedicated dataset of a polarizing topic, Brexit of around 770 Article Parts, to see whether the models captured the polarizing and mostly negative responses to that topic.

However due to an oversight as well as later resource and time constraints, these Articles were all written in the U.S., this does not change the evaluation however it is important to note, as it shows clear preference to U.S. Figures.

"True_Data" (GPT):

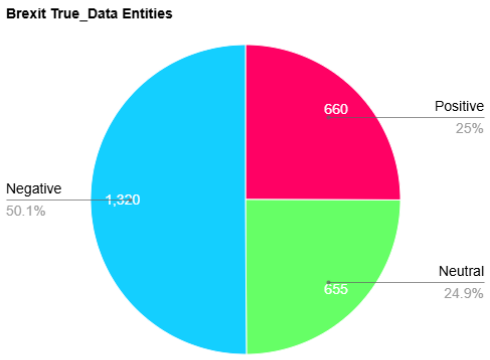

Figure 12 Brexit Overall Attitude showcase True Data Entities

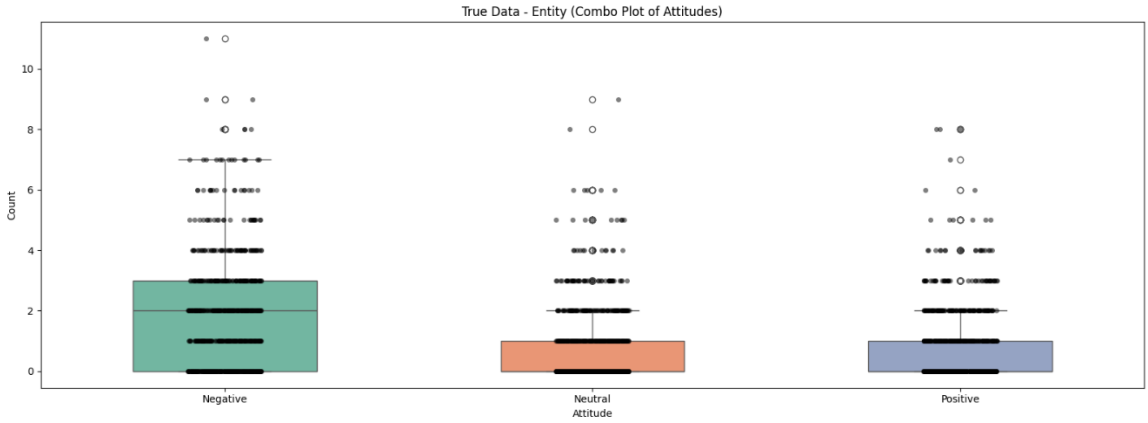

Figure 13 Brexit Overall Attitude Seaborn Graph True Data Entities

These Graphs show us that for the Brexit dataset, Entity Pairs, GPT covers a majority of Negative pairings, with some Positives/Neutral, averaging 0-1 Positive/Neutral and 0-3 Negative assigns per Sample, with the highest being 11 Negative assigns for 1 sample.

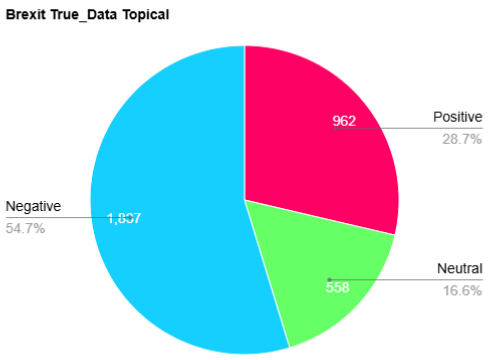

Figure 14 Brexit Overall Attitude showcase True Data Topical

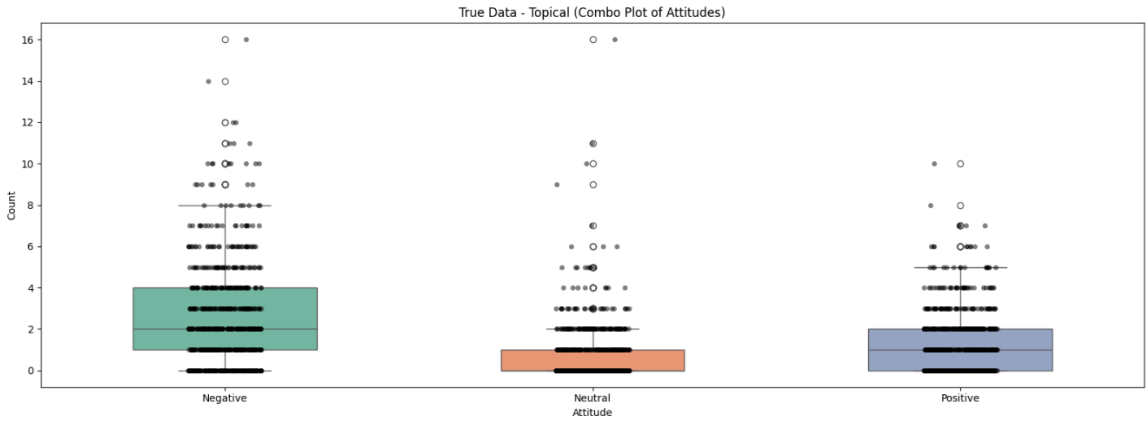

Figure 15 Brexit Overall Attitude Seaborn Graph True Data Topical

These Graphs show us that for Topical Pairs, GPT covers a larger majority of Negative pairings, with some Positives/Neutral, averaging 0-1 Neutral, 0-2 Positive and 1-4 Negative assigns per Sample, with the highest being 16 Negative assigns for 1 sample.

"Pred_data" (POLAR):

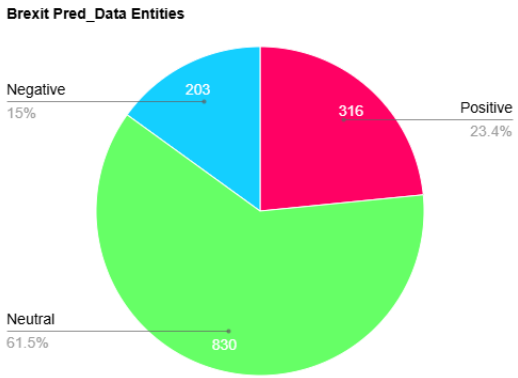

Figure 16 Brexit Overall Attitude showcase Predicted Data Entities

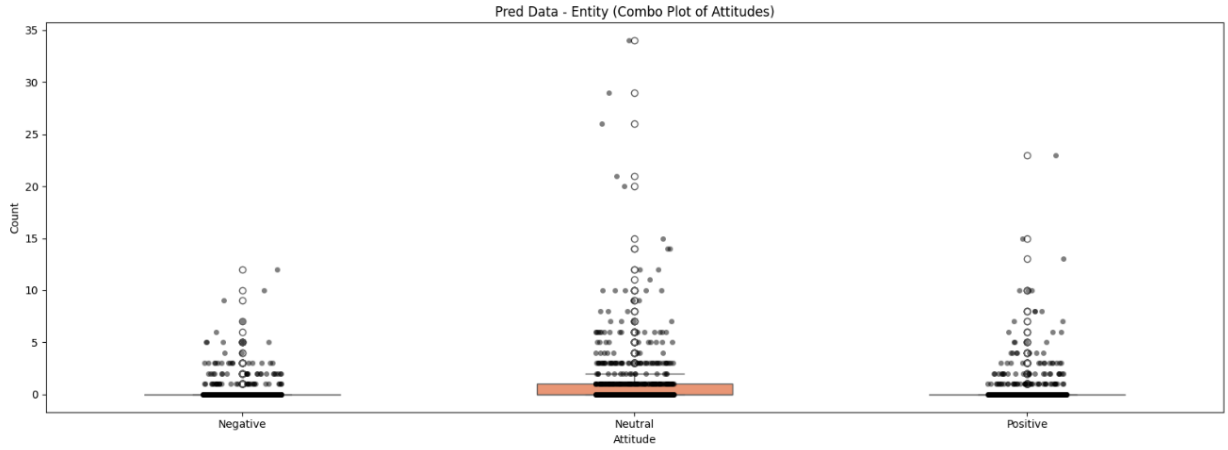

Figure 17 Brexit Overall Attitude Seaborn Graph Predicted Data Entities

These Graphs show us that for Entity Pairs, POLAR covers a majority of Neutral pairings, with some Positives/Negative, averaging 0-1 Neutral and ≈0 Negative/Positive assigns per Sample, with the highest being 34 Negative assigns for 1 sample.

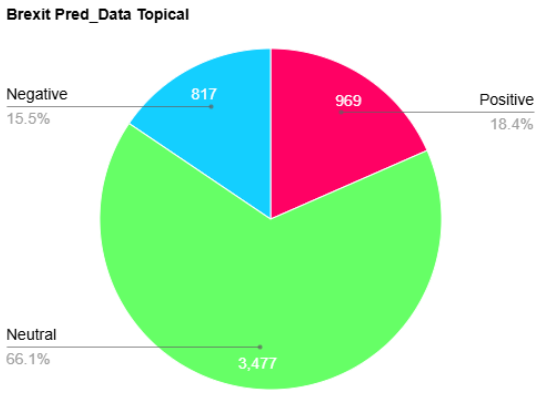

Figure 18 Brexit Overall Attitude showcase Predicted Data Topical

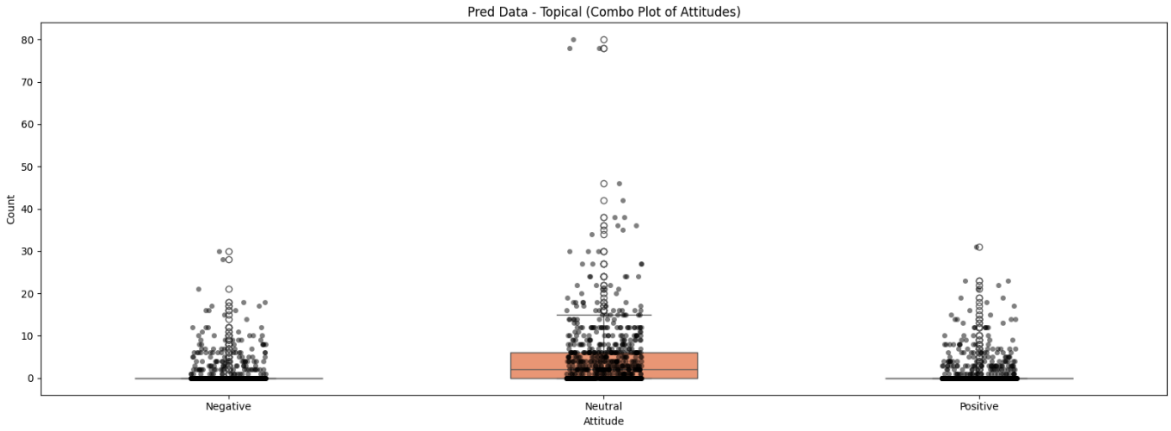

Figure 19 Brexit Overall Attitude Seaborn Graph Predicted Data Topical

These Graphs show us that for Topical Pairs, POLAR covers a majority of Neutral pairings, with some Positives/Negative, averaging 0-7 Neutral and ≈0 Negative/Positive assigns per Sample, with the highest being 80 Negative assigns for 1 sample.

6.2.2.2 Overall Attitude Summary Conclusion

The results in this section highlight a clear divergence in attitude assignment between GPT and POLAR, across both datasets a consistent pattern was shown:

- GPT distributes Positive/Neutral/Negative attitudes with high granularity and diversity, adapting its usage to the tone and context of each sample.
- POLAR, instead, exhibits a dominant bias toward Neutrality, frequently assigning Neutral to pairs, whereas GPT infers more polarized relationships

This is shown due to the structural limitations of POLAR, being error-prone to dependency parsing errors as well as sentiment scoring through pre-existing lexicons, which fail at more nuanced or complex sentences.

Training Dataset:

In the General training dataset comprising of multiple non-organized likely polarizing topics.

- GPT assigned Positive/Neutral/Negative ratings in an organized fashion that reflects general diversity between articles, such as praise, criticism, disagreement etc.
- POLAR however assigned Neutral to the overwhelming majority of both Entity and Topical pairs, with the Average Positive/Negative rating per Article averaging at ≈0 for both, whereas the Neutral ratings averaging 0-1 to 0-7 instead.

Brexit Dataset:

The Brexit dataset served as a biased test case, since Brexit related discourse is inherently polarizing and often emotionally charged, leading people to feel strongly about either side of polarity.

- GPT's responses reflect this polarity, assigning a strong majority of the pairs as Negative, for both Entities and Topics, followed by a smaller majority of them being Positive for Topics and even with Neutral assignments for Entities.
- POLAR, on the other hand, continued its overwhelming Neutral dominant pattern, with no clear sign of preference to Negative attitudes unlike GPT's responses. This is likely due to dependency errors or a lack of coverage by the sentiment lexicon (mpqa)

This shows that POLAR tends to underestimate Negativity in politically charged datasets, which impacts its general applicability to stance detection for Entities, mapping controversies etc.

Either this or POLAR failed to find any emotionally intensive Pairs or polarizing relationships within this dataset, however this is proven false from Section 6.2.4

6.2.3 Attitude Analysis Conclusion

These results demonstrate that while POLAR's Sentiment Extraction provides a structured and interpretable approach, it can drastically underrepresent polarity in real world articles.

LLMs like GPT and Mistral can significantly improve on this aspect of Sentiment Extraction, offering a richer and more descriptive Attitude analysis as well as context-awareness when evaluating Attitudes.

6.2.4 Topics and Pair Frequency Comparisons

To analyze a broader range of comparisons between POLAR and GPT, the Topics extracted through the Pipeline were compared, which is an independent step from the Sentiment Attitudes. We also analyzed the Pair Frequency of the Brexit Dataset for both POLAR and GPT to see through a broader scope whether the models identified similar pairs and their calculated attitudes.

6.2.4.1 Topics

In this section we compared the Topics file generated by the Topic_Identifier step of the POLAR Pipeline, see Chapter 2, Section 2.1.2.4, this was done by treating the Topics in the Topical Pairs as Noun_Phrases and reformatting them to generate the topics.json.gz.

We compared the LLM integrated topics file with the normal POLAR topics file.

These Comparisons were done on the Brexit Dataset.

Recall	*0.7294*
True Positives	*2235*
False Negatives	*829*
POLAR Topics	*2671*
GPT Topics	*2628*
Unmatched POLAR	*436*
Unmatched GPT	*393*

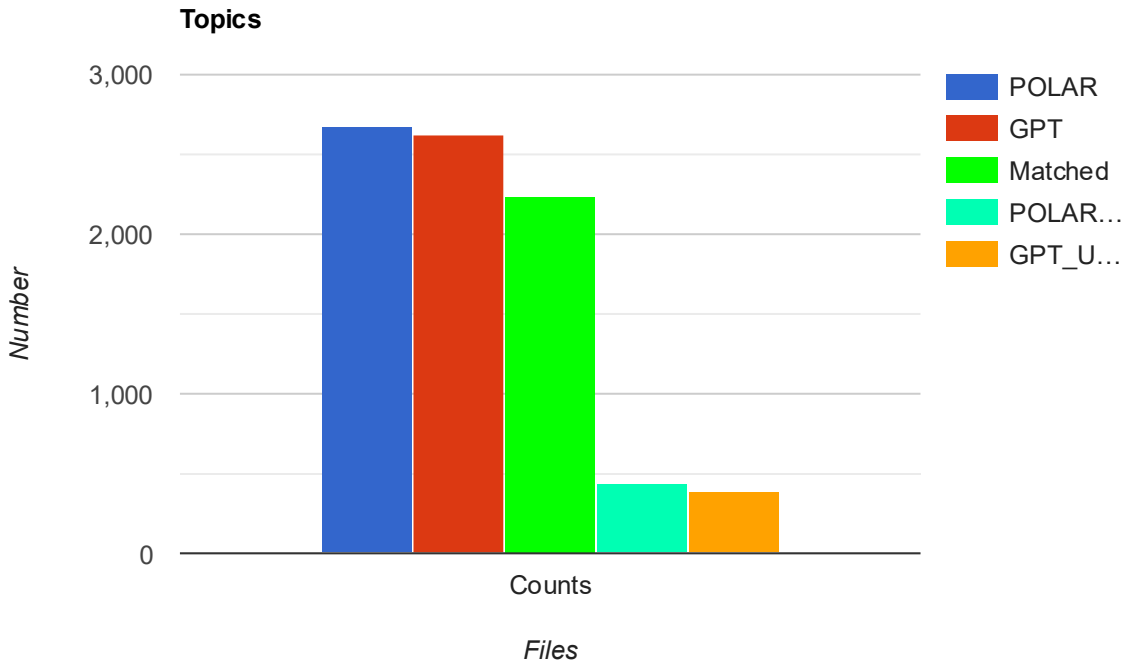

Figure 20 Brexit Similar Topics between models

From this graph we can see that, even though the per-article analysis proved that these Datasets are wildly different, their calculated Topics were very similar, where 83.68% of the POLAR Attitudes were matched with GPT Attitudes and 85.05% of GPT Attitudes were matched with POLAR Attitudes.

These results are aligned with POLAR's Topic Calculations done in the POLAR Journey, with very similar results to the ones gathered from that research. **[1]**

Table 8. Results for the Topic Identification Accuracy

Case Study	Our Framework	PaCTE	LOE
Abortion	0.88	**1.00**	**1.00**
Immigration	0.60	**0.80**	**0.80**
Gun Control	**0.83**	**0.83**	**0.83**
Average	0.77	**0.87**	**0.87**

The Topics were compared to each other using Semantic Similarity with Sentence Transformers with a Threshold of 0.6.

Examples of matches with the lowest Threshold:
"a potential Brexit" <-> "The so-called 'Brexit' decision" , score: 0.6003
"a political crisis" <-> "economic and political uncertainty" , score: 0.609

6.2.4.2 Pair Frequency

Utilizing the results gathered from comparing each individual Sample, we gathered every unique mention of a Pair in every Sample for both Entity/Topical fields and counted every time each Pair is mentioned in both the Predicted Data and the Ground Truth Data.

We grouped up every mention of each pair as well as every time the pair was mentioned in both the Ground Truth set and the Predicted set, as well as every time it was declared as Positive, Neutral or Negative.

It is important to note that this step was not done using Semantic Similarity to compare the pairs due to Time Constraints, since to accurately and fairly compare every pair with every pair using Semantic Similarity would take multiple days of execution, however would likely yield better results.

Overlapping Entity Pairs:

Overlapping Entity Pairs (113) == Top 10

Where True Attitudes refer to GPT and Predicted Attitudes refer to POLAR:

Pairs	***True_Attitudes***	***Pred_Attitudes***
('european union', 'united kingdom')	*TRUE: 72 (Neg: 47, Neu: 18, Pos: 7)*	*PRED: 109 (Neg: 14, Neu: 80, Pos: 15)*
('brexit', 'european union')	*TRUE: 7 (Neg: 7, Neu: 0, Pos: 0)*	*PRED: 44 (Neg: 10, Neu: 24, Pos: 10)*
('brexit', 'united kingdom')	*TRUE: 5 (Neg: 1, Neu: 2, Pos: 2)*	*PRED: 43 (Neg: 9, Neu: 28, Pos: 6)*
('united kingdom', 'united states')	*TRUE: 29 (Neg: 5, Neu: 6, Pos: 18)*	*PRED: 6 (Neg: 0, Neu: 5, Pos: 1)*
('barack obama', 'european union')	*TRUE: 11 (Neg: 1, Neu: 4, Pos: 6)*	*PRED: 13 (Neg: 4, Neu: 6, Pos: 3)*
('barack obama', 'united kingdom')	*TRUE: 16 (Neg: 5, Neu: 4, Pos: 7)*	*PRED: 8 (Neg: 1, Neu: 4, Pos: 3)*
('donald trump', 'hillary clinton')	*TRUE: 21 (Neg: 19, Neu: 0, Pos: 2)*	*PRED: 1 (Neg: 0, Neu: 1, Pos: 0)*
('brexit', 'donald trump')	*TRUE: 13 (Neg: 5, Neu: 2, Pos: 6)*	*PRED: 7 (Neg: 0, Neu: 6, Pos: 1)*
('nato', 'russia')	*TRUE: 15 (Neg: 11, Neu: 2, Pos: 2)*	*PRED: 3 (Neg: 0, Neu: 2, Pos: 1)*
('barack obama', 'brexit')	*TRUE: 6 (Neg: 3, Neu: 2, Pos: 1)*	*PRED: 10 (Neg: 3, Neu: 4, Pos: 3)*

Given the topic of Brexit and the general timeframe of the articles of it happening 2016/6-7/22-10, (European Union <-> United Kingdom) is the most represented pair and that lines up with the topic perfectly, with 72 mentions from GPT and 109 from Polar, with the same reasoning for the 2nd and 3rd row, with a small oversight being that Brexit should have been considered a topic rather than an Entity.

The Entities that are matched are accurate to the Topic of Brexit, however the Attitudes associated with them are as mentioned in Section 6.2.2.2. POLAR's Attitudes are primarily Neutral, even with the predominantly Negative nature of the Pairs, such as: (European Union <-> United Kingdom), should be the most polarizing pair, given real world scenarios of what happened, and GPT encapsulates that by having 65.27% of these pairs be Negative, followed by 25% Neutral and only 9.7% Positive. Meanwhile POLAR has 73.39% of these pairs be classified as Neutral, with 13.76% Positive and only 12.84% Negative.

Regardless of whether these pairs were found in the same Sample or not, they were found in the same Dataset, so the pairs being Neutral focused is due to POLAR rather than the dataset, since GPT attributed a majority of them as Negative.

Overlapping Topical Pairs:

Overlapping Topical Pairs (45) == Top 10

Where True Attitudes refer to GPT and Predicted Attitudes refer to POLAR:

Pairs	***True_Attitudes***	***Pred_Attitudes***
('brexit', 'european union')	*TRUE: 70 (Neg: 58, Neu: 10, Pos: 2)*	*PRED: 20 (Neg: 6, Neu: 14, Pos: 0)*
('european union', 'united kingdom')	*TRUE: 12 (Neg: 7, Neu: 5, Pos: 0)*	*PRED: 47 (Neg: 6, Neu: 34, Pos: 7)*
('brexit', 'united kingdom')	*TRUE: 34 (Neg: 22, Neu: 9, Pos: 3)*	*PRED: 21 (Neg: 2, Neu: 16, Pos: 3)*
('barack obama', 'brexit')	*TRUE: 24 (Neg: 14, Neu: 8, Pos: 2)*	*PRED: 1 (Neg: 0, Neu: 1, Pos: 0)*
('brexit', 'united states')	*TRUE: 15 (Neg: 8, Neu: 4, Pos: 3)*	*PRED: 9 (Neg: 1, Neu: 4, Pos: 4)*
('european union', 'united states')	*TRUE: 4 (Neg: 1, Neu: 0, Pos: 3)*	*PRED: 5 (Neg: 1, Neu: 4, Pos: 0)*

('boris johnson', 'brexit')	*TRUE: 8 (Neg: 1, Neu: 0, Pos: 7)*	*PRED: 1 (Neg: 0, Neu: 1, Pos: 0)*
('brexit', 'david cameron')	*TRUE: 8 (Neg: 7, Neu: 1, Pos: 0)*	*PRED: 1 (Neg: 0, Neu: 1, Pos: 0)*
('barack obama', 'european union')	*TRUE: 5 (Neg: 1, Neu: 2, Pos: 2)*	*PRED: 3 (Neg: 1, Neu: 1, Pos: 1)*
('united kingdom', 'united states')	*TRUE: 2 (Neg: 1, Neu: 1, Pos: 0)*	*PRED: 5 (Neg: 1, Neu: 3, Pos: 1)*

The Topical pairs are once again accurate to the events of Brexit, the Outliers are the pairs consisting of 2 entities instead of an entity and a topic, example:
('european union', 'united kingdom')
('united kingdom', 'united states')

These same pairs are found above in the Entity Pairs

6.2.4.3 Entity Frequency

Apart from looking at the most represented pairs from each model, it is also important to look at the Entity Frequencies of each mention, to further see the similarity between the models.

These are all the Mentions of Entities, disregarding whether the pairs they were calculated with, to see whether the models truly extracted similar data in terms of the Entities calculated, the Entities could either be Entity1/2 in any pair of the Entity Pairings.

Top 10 most represented of each model:

Top 10 Entity Mentions in the **True Dataset (GPT)**

Entity	***Mentions***
United Kingdom	Count: 343 (Pos: 97, Neu: 78, Neg: 168)
European Union	Count: 332 (Pos: 78, Neu: 61, Neg: 193)
Brexit	Count: 239 (Pos: 41, Neu: 34, Neg: 164)
Donald Trump	Count: 210 (Pos: 39, Neu: 57, Neg: 114)
Barack Obama	Count: 123 (Pos: 41, Neu: 37, Neg: 45)
United States	Count: 95 (Pos: 38, Neu: 21, Neg: 36)
David Cameron	Count: 94 (Pos: 15, Neu: 16, Neg: 63)
Boris Johnson	Count: 67 (Pos: 6, Neu: 20, Neg: 41)
Nigel Farage	Count: 63 (Pos: 6, Neu: 18, Neg: 39)
Russia	Count: 60 (Pos: 16, Neu: 9, Neg: 35)

Top 10 Entity Mentions in the **Predicted Dataset (POLAR)**

Entity	***Mentions***
United Kingdom	Count: 331 (Pos: 66, Neu: 211, Neg: 54)
Union	Count: 291 (Pos: 61, Neu: 180, Neg: 50)
Brexit	Count: 230 (Pos: 52, Neu: 121, Neg: 57)
European Union	Count: 121 (Pos: 27, Neu: 81, Neg: 13)
London	Count: 59 (Pos: 9, Neu: 38, Neg: 12)
Barack Obama	Count: 55 (Pos: 15, Neu: 28, Neg: 12)
Reuter	Count: 50 (Pos: 9, Neu: 39, Neg: 2)
Donald Trump	Count: 46 (Pos: 10, Neu: 32, Neg: 4)
Warsaw	Count: 38 (Pos: 11, Neu: 25, Neg: 2)
Brussel	Count: 35 (Pos: 19, Neu: 15, Neg: 1)

In terms of Highest Frequency of Entities, the models correctly identify the same top 3 most represented Entities, although with differing Attitudes assigned to them, this is done however by POLAR treating Union and European Union as a different entity, thus the European Union would be the most represented, which also makes sense given the topic. The other entities apart from Barack Obama and Donald Trump also differ later.

This table data shows us that even though from previous results in 6.2.1.1, GPT tends to detect more Entity Relationships than POLAR, the large majority of POLAR's pairs tend to include the most prominent Entities.

6.3 POLAR Pipeline End Result Comparison

In this section we are moving past the Sentiment Attitude Extraction aspect of POLAR and looking at the Artifacts output at different Steps of the Pipeline and comparing the Standard POLAR method with our Integrated LLM Method.

6.3.1 Fellowships:

POLAR:


```
{"fellowships": [[" United_States", " London", " Reuters", " Paris", " Nigel_Farage", " Great_Britain", " Ringo_Starr", "
George_Osborne", " Annexation_of_Crimea_by_the_Russian_Federation"], [" European_Union", " NATO", " Russia", " Ukraine"],
[" Brexit", " Vladimir_Putin", " Brussels", " European_Council"], [" California_Republican_Party", " Hillary_Clinton", "
Bernie_Sanders"], [" Demography_of_the_United_Kingdom", " President_of_the_United_States", "
Government_of_the_United_Kingdom"], [" United_Kingdom", " Acts_of_Union_1800", "
Speaker_of_the_United_States_House_of_Representatives"], [" England", " Wales"], [" Barack_Obama", " Warsaw"], [" Texas",
" Texas_secession_movements"], [" Conservative_Party_(UK)"], [" Federal_government_of_the_United_States"], ["
Fran\u00e7ois_Hollande"], [" President_of_France"], [" David_Cameron"], [" Steve_Hilton"], [" Boris_Johnson"], ["
Scotland"], [" Refugees_of_the_Syrian_civil_war"], [" British_Empire"], [" Hafiz_Saeed"], [" Interpol_notice"], ["
Base_erosion_and_profit_shifting"], [" United_Nations"], [" Italy"], [" Jeremy_Corbyn"], [" Scottish_Labour"], ["
BBC_News"], [" France"], [" Bloomberg_News"], [" George_Washington"], [" Elizabeth_II"], [" African_Americans"], ["
Hispanic_and_Latino_Americans"], [" Bauer_Media_Audio_UK"], [" Bauer_Media_Group"], [" American_Automobile_Association"],
[" Mary_Robinson"], [" Parliament_of_the_United_Kingdom"], [" Scottish_National_Party"], [" Dallas"], ["
The_Early_Show"], [" Donald_Trump"]]}
```


Figure 21 POLAR's Fellowship Structure

GPT:

```
{"fellowships": [["European Union", "Barack Obama", "NATO"], ["EU", "Germany", "Russia"], ["Britain", "United Kingdom",
"United States"], ["Brexit", "Donald Trump"], ["UK", "Northern Ireland"], ["President Barack Obama", "Prime Minister
David Cameron"], ["China", "George Osborne"], ["Eric Schmidt", "Google"], ["Boris Johnson"], ["David Cameron"], ["Sadiq
Khan"], ["President Obama"], ["Nigel Farage"], ["Institution"], ["Wright"], ["U.K."], ["John McLaren"], ["Tim North"],
["Canada"], ["Leave campaign"], ["Remain campaign"], ["Michael Gove"], ["Olli Rehn"], ["Peter Sutherland"], ["France"],
["Ruth Davidson"], ["Jacqueline Towers-Perkins"], ["Dave Schick"], ["Fauci"], ["Bernie Sanders"], ["Hillary Clinton"],
["Clinton"], ["Trump"], ["Brexit supporters"], ["Tony Blair"], ["Republican establishment"], ["Trump supporters"],
["David Herz"], ["Turkey"], ["Brexit referendum"], ["Scotland"], ["Harvard economist Carmen Reinhart"], ["Robert Margo"],
["Tom Cotton"], ["Natalie Nougayrede"], ["Peter Westmacott"], ["Richard Haass"], ["GOP electorate"], ["Paul Ryan"],
["Switzerland"], ["Great Britain"], ["Donald Tusk"], ["Republic of Ireland"], ["Nicola Sturgeon"], ["Angela Merkel"],
["US"], ["Geert Wilders"], ["Poland"], ["The United States"], ["BJP"], ["liberals"], ["William Hague"], ["The European
Union"], ["UK Independence Party (UKIP)"], ["Older people"], ["Young people"], ["England"], ["London"], ["Areeq
Chowdhury"], ["UK Independence Party"], ["Zack Beauchamp"], ["J.K. Rowling"], ["Leave movement"], ["The Germans"], ["The
Greeks"], ["The British"], ["The EU"], ["Sonia Purnell"], ["Leave voters"], ["Yascha Mounk"], ["Viktor Orban"], ["Gwynne
Dyer"], ["Jeremy Corbyn"], ["Labour Party"], ["Goldman Sachs"], ["EU leaders"], ["John Kerry"], ["American
conservatives"], ["Josh Marshall"], ["British Empire"], ["British people"], ["Editor"], ["Ukraine"], ["Republicans"],
["Philip Morris International"], ["Imperial Tobacco"], ["Japan Tobacco International"], ["Richard Branson"], ["Institute
of Directors"], ["Fran\u00e7ois Hollande"], ["Marine Le Pen"], ["Americans"], ["French Foreign Office"], ["EC"],
["French"], ["Mr. Renzi"], ["Voters in Great Britain"], ["Justin Trudeau"], ["Martin Schulz"], ["Leave vote"], ["Richard
Balfour-Lynn"], ["Iceland's prime minister"], ["John Kasich"], ["Mr. Jackson"], ["India"], ["Morgan Stanley"], ["British
voters"], ["anti-EU warriors"], ["Europe"], ["Doukhobors"], ["OnePlus"], ["Europeans"], ["Parents of the child"],
["Teesville Primary School staff"], ["Ringo Starr"], ["Leavers"], ["Alyn Smith"], ["David Edward"], ["Tim Peake"],
["Arundhati Bhattacharya"], ["SBI"]]}
```


Figure 22 GPT's Fellowship Structure

From these Fellowships we can infer multiple things:

GPT:

- 110+ Clusters of Fellowships, ranging from 1-3 Entities each with a majority of single Entities, many overlaps (European Union – EU, etc.)
- Contains Various named Entities not limited to strictly known entities
- High Noise, a lot of Single Entity Fellowships

POLAR:

- ~40 Clusters of Fellowships, ranging from 1-9 Entities each, more compact with less overlaps
- Represents Entities Strictly off Dbpedia links that correspond to Entities
- Very low Noise, very curated

Very few similarities between the 2 models, with the closest match in fellowships being:

"POLAR ": ["Demography_of_the_United_Kingdom",
"Government_of_the_United_Kingdom"
"President_of_the_United_States"],

"GPT ": ["Britain",
"United Kingdom",
"United States"]

Comparing the fellowship clusters of POLAR and GPT reveals two distinct modeling philosophies.

POLAR's DBpedia-aligned clusters are compact and curated, offering high precision but missing some aspects of political and sociological discourse.

GPT, on the other hand, surfaces a richer, more sociologically nuanced set of actors, including campaign groups, media personalities, and ideologically charged with the trade-off of a messier and noisier output.

6.3.2 SAG (Signed Attitude Graph):

In this section we compared the SAG graphs output by POLAR's pipeline which shows the Sentiment between Entities in an undirected graph.

What the Connections represent:

The Nodes represent Entities and the Edges represent Sentiment calculated between Entities whose pairs were found through the Entity_Extractor and Sentiment_Attitude, or in the case of the LLMs, the Entity_attitude field from the responses.

The Colors represent whether the weight attached to these pairs is Positive or Negative. The POLAR function in the pipeline that constructs the SAG skips over strictly Neutral pairs, to focus more on Polarization.

POLAR:

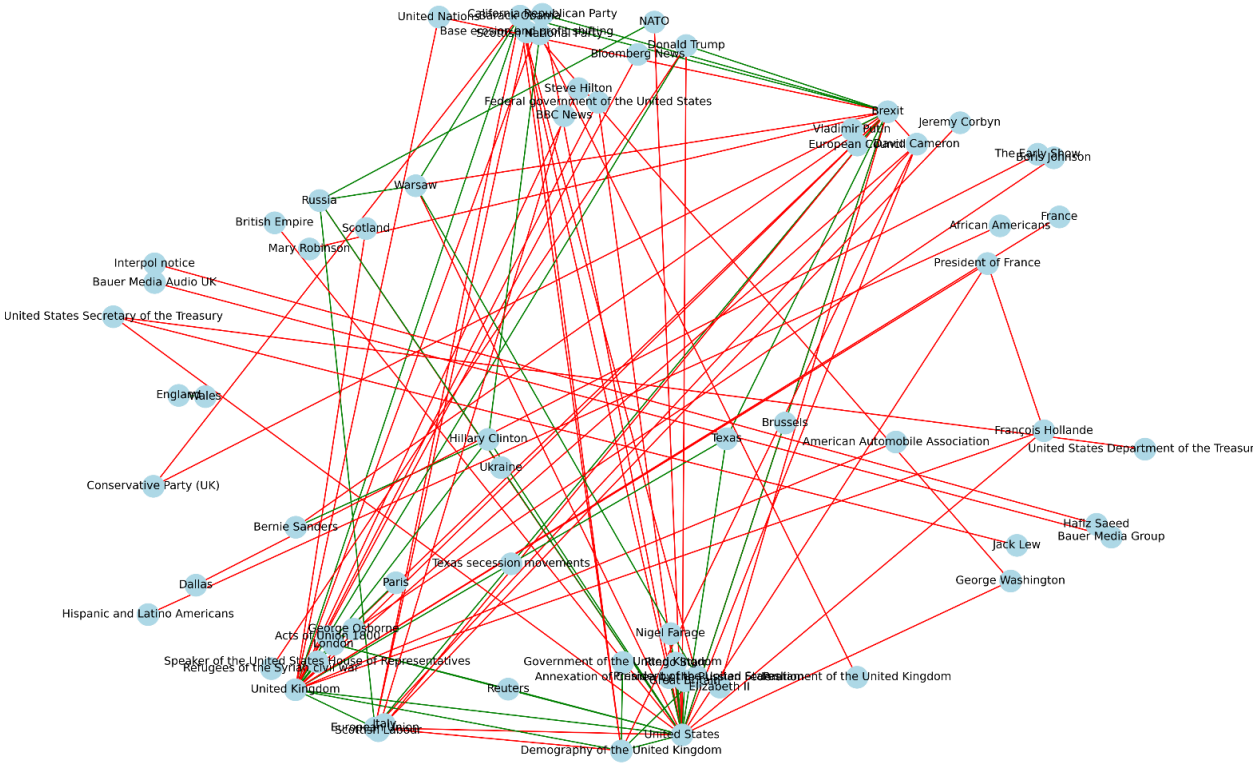

Figure 23 POLAR's Signed Attitude Graph (SAG)

GPT:

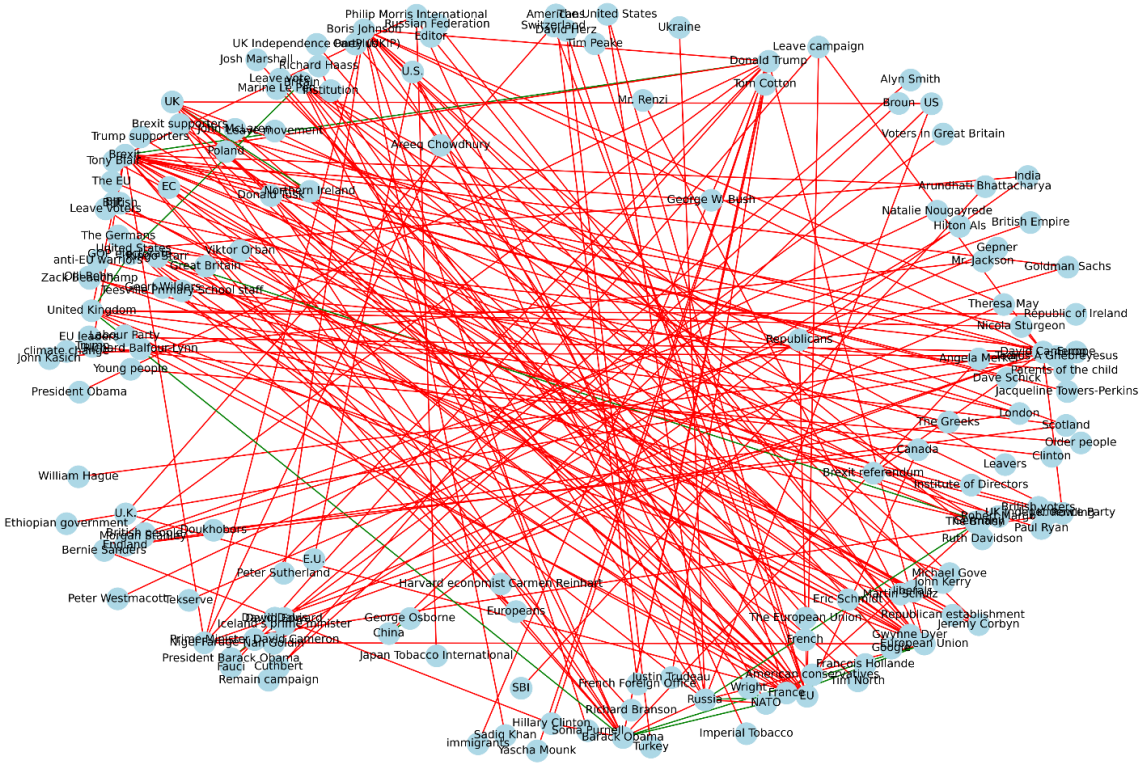

Figure 24 GPT's Signed Attitude Graph (SAG)

From these Graphs we can see:

POLAR

- SAG is more compact and structured, containing less noise
- Entities are Formal and linked to dbpedia URLs each (removed from the Graph for clarity)
- Mix of Positive and Negative edges, leaning heavily towards Negative

GPT

- Messy Graph with a high number of Nodes and Edges, very densely connected
- Heavily skewed towards Negative Sentiment Edges with only a handful of Positives
- Entities are informal mentions through text, not linked to pre-existing Entities

These results are very dissimilar, likely because GPT overproduces Entity Pairs, as seen in section 6.2.2.1.2, POLAR produces 1320 Entity Pairs to GPT's 2552. Added in the fact that GPT's Entity Pairs are not traditionally normalized like Polar's use of Dictionary linking to dbpedia links, and this creates this difference we see, with GPT appearing messier and denser, with noisy Entities that appear few times.

6.3.3 Attitude Dipoles:

This is the final Artifact output from the POLAR Pipeline, the attitudes.pckl file, which is the culmination of all the previous steps in the Pipeline, it shows all of the key entities' connection with the other key entities by linking them with their opinions on the most impactful topics found.

What each Color Represents:

Blue: Neutral to slight Agreement on the Topic
Gray: Mixed to Neutral on the Topic
Green: Positive Agreement on the Topic
Orange: Negative Agreement on the Topic
Red: Polarization on the Topic

POLAR:

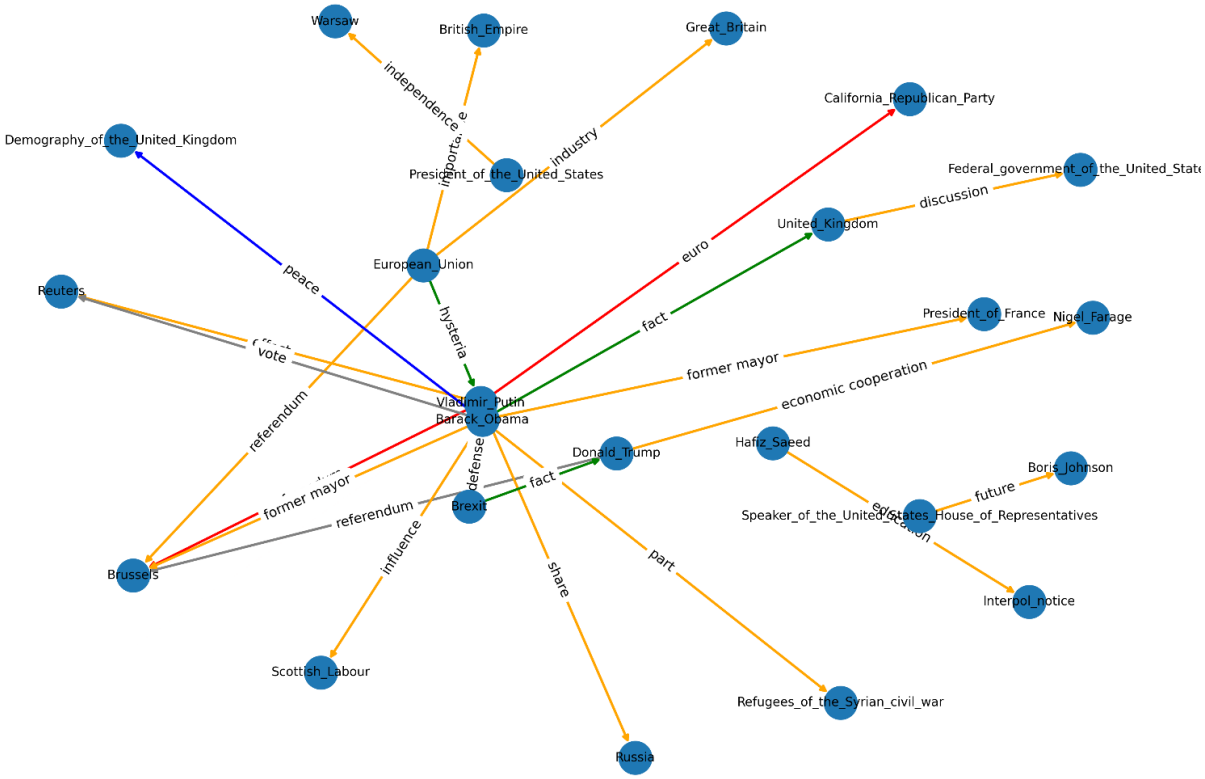

Figure 25 POLAR Dipole Attitude Graph

GPT:

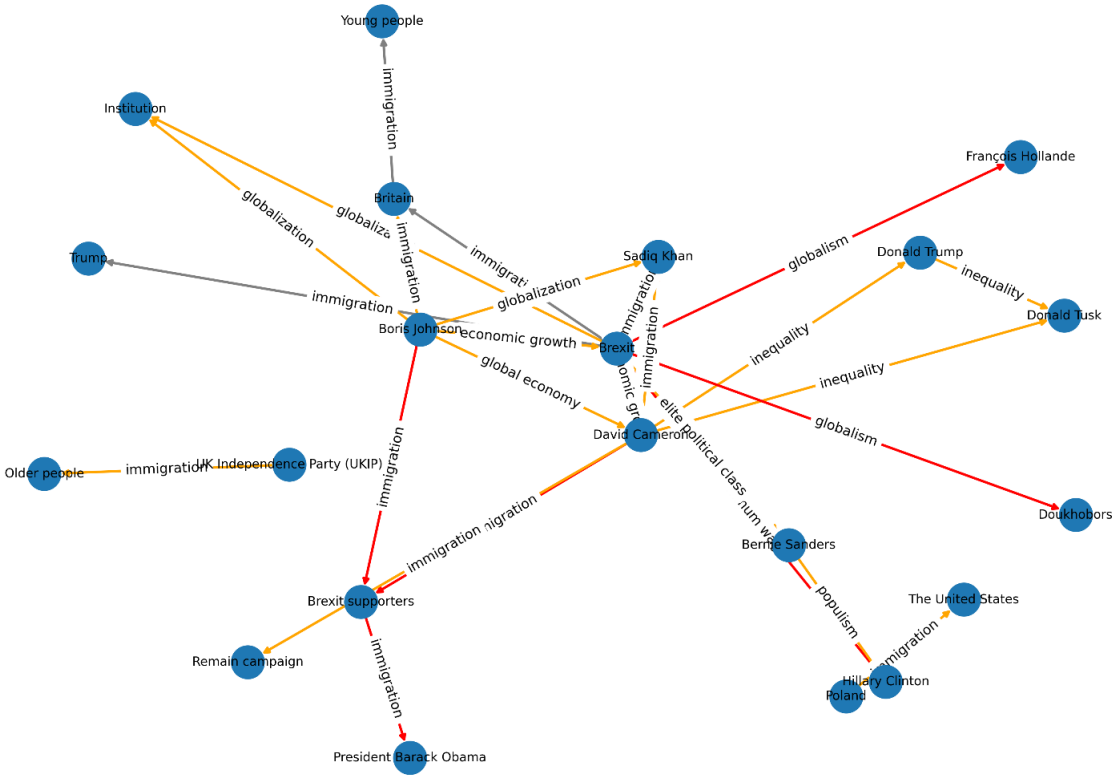

Figure 26 GPT Dipole Attitude Graph

Analysis:

Initially, at a first glance the models' outputs share some Similarities, such as having a large majority of the Edges being Negative Agreement, where both Entities Disagree on the Topic of their Pair, such as:

- GPT (David Cameron <-> Donald Trump, Topic: inequality)
- POLAR (Barack_Obama <-> Russia, Topic: share).

The number of Polarizing Pairs are few for POLAR and a bit more prevalent in GPT's graph, with pairs such as:

- GPT (Boris Johnson <-> Brexit supporters, Topic: immigration, where Boris Johnson: Positive Attitude, Brexit supporters: Negative Attitude)

- POLAR (Vladimir_Putin <-> California_Republican_Party, Topic: euro, where Vladimir_Putin: Negative Attitude, California_Republican_Party: Positive Attitude)

There are slight Positive Agreements and a few Neutral Pairs as well.

The Topics that connect the pairs in the models vary substantially for GPT and POLAR, where POLAR's Topics are plentiful and nuanced, but do not seem to be Topics of high Polarity or Discourse, consisting mainly of generic Noun Phrases, such as:

- Part
- Share
- Fact
- Discussion
- Former mayor
- Future
- euro

On the other hand, GPT's Topics are very few comparatively, with only 17 unique Topics found in the Dipoles out of 43, with immigration appearing in 16 entries (roughly 37%). As such the Topic coverage of GPT is less nuanced and consists of some repeating topics between Entities, however the Topics being compared are divisive and are likely to cause Discourse, such as:

- immigration
- globalism
- globalization
- inequality
- global economy

In order to determine the quality of the Topics gathered by GPT and POLAR, the Topics were all sent to multiple LLMs for them to classify as Sociopolitical, Slightly Sociopolitical and Not Sociopolitical, with a brief explanation that Sociopolitical topics are issues related to government, society, public policy, inequality, rights, identity or power structures and may be emotionally charged in public discourse.

The topics were sent randomly with no reference to which response they correlate to, in order to remove any inherent bias in the outputs, with these results gathered from purely considering the Sociopolitical classified topics:

Sociopolitical Topics detected using LLMs

Model	*GPT Percentage of Matches*	*POLAR Percentage of Matches*
GPT-4.0	*100%*	*45.1%*
Deepseek-V3	*76.7%*	*40.7%*
Claude-3.7 Sonnet	*74.4%*	*29.7%*
Gemini-2.0 Flash	*76.7%*	*44.0%*
Le Chat - Mistral	*72.1%*	*31.9%*
Average Overall:	***80.0%***	***38.3%***

6.3.4 Conclusion:

It is hard to definitively declare which model is superior in the end, a more thorough analysis would need to be conducted with a much larger dataset of Articles to be able to more accurately compare the models.

From the Data of 727 Article Parts however, on the topic of Brexit, the analysis between GPT and POLAR reveals a clear trade-off between topic depth and topic breadth:

GPT overwhelmingly identifies socio-politically meaningful topics, as validated by multiple LLMs. Focusing on issues with real-world ideological divides.

In contrast, POLAR extracts a much larger and diverse topic set of 65 unique topics of 91 total Attitudes to GPT's 17 of 43 total Attitudes, as well as reflecting a broader topical coverage when compared to GPT. The trade-off is that many of POLAR's topics are shallow or generic, with no means of discourse or polarity, with only a small portion of sociopolitical topics covered (roughly 38.3% according to the LLMs requested)

6.4 Fine-Tuned Mistral Evaluation

This evaluation aims to measure how well the fine-tuned Mistral-7B model replicates the outputs of GPT3.5. The GPT responses were set as the Ground Truth, with the Mistral responses acting as the prediction, the comparisons were focused on entity/topic pair identification, attitude classification on the true pairs and justification quality.

It is important to note that the Dataset that was compared was a dataset excluded from the fine-tuning dataset of the model.

6.4.1 Evaluation

Evaluation - Mistral completed in 4467.64 seconds.(~74.46 minutes)

"Entity"

Pair Matching

Precision	*0.8725*
Recall	*0.9418*
F1	*0.9059*
True Positives	*664*
False Positives	*97*
False Negatives	*41*

Attitude Prediction

Accuracy	*0.9771*
True Positives	*683*
False Negatives	*16*

Justification Overlap

Overlap	*0.8581*
Matched Justifications	*798*
Total Justifications	*930*

From this data, we can see that from the evaluated dataset, Mistral Successfully manages to predict 664 pairs from the 705 pairs in the Ground Truth dataset == roughly 94.18% of the pairs in Ground Truth were accurately predicted.

Of those 664 pairs, 699 attitudes were compared between the Ground Truth and the Predicted Dataset, with Mistral correctly predicting 683 of them == roughly 97.71% pairs were correctly attributed.

Of those 664 pairs, the justification was successfully matched using ROUGE-L similarity roughly 85.81% of the time.

“Topical”:

Pair Matching

Precision	*0.8658*
Recall	*0.9365*
F1	*0.8998*
True Positives	*826*
False Positives	*128*
False Negatives	*56*

Attitude Prediction

Accuracy	*0.9750*
True Positives	*820*
False Negatives	*21*

Justification Overlap

Overlap	*0.8785*
Matched Justifications	*477*
Total Justifications	*543*

From this data, we can see that from the evaluated dataset, Mistral Successfully manages to predict 826 pairs from the 882 pairs in the Ground Truth dataset == roughly 93.65% of the pairs in Ground Truth were accurately predicted.

Of those 826 pairs, 841 attitudes were compared between the Ground Truth and the Predicted Dataset, with Mistral correctly predicting 820 of them == roughly 97.50% pairs were correctly attributed.
Of those 826 pairs, the justification was successfully matched using ROUGE-L similarity roughly 87.85% of the time.

6.4.2 Conclusion

From this dataset we can confidently state that Mistral replicates GPT remarkably well, accurately guessing GPT's Entity / Topical Pairs around 93-94% of the time, with a precision of around 86-87%, meaning that at any point that Mistral locates a pair, there is an 86-87% likelihood that it is included in the GPT results. These are incredible results, especially since Mistral is a marginally smaller open-weight model than GPT 3.5.

Mistral is also incredibly accurate at predicting the Attitudes of the calculated pairs, accurately guessing around 97% of the pairs in the same way that GPT would.

These results indicate that Mistral, once fine-tuned, is not only a viable replacement for GPT-based processing in the POLAR pipeline, but also one that can produce slightly richer outputs. In the test set, an additional 97 Entity Pairs and 128 Topical Pairs were identified by Mistral that do not match any pairs in the GPT outputs. This can be interpreted as an enhancement (greater granularity and sensitivity to subtle Sentiment Attitudes) or as potential noise (false pairs with lower precision). The appropriate interpretation of these added pairs depends on the strictness or flexibility of the use case.

Beyond raw performance metrics, there are operational trade-offs that distinguish the two models:

Attribute	***GPT-3.5 (API)***	***Mistral-7B (Fine-Tuned)***
Hosting	*Cloud (OpenAI)*	*Local*
Internet Dependency	*Required*	*None*
Prompt Length	*Long (1200+ char)*	*Minimal*
Latency per Request	*~15–35 seconds*	*~1–2 minutes*
Multithread-ability	*Minimal*	*None*
Cost per Use	*Usage-based fees*	*Local*
Fine-tune Time	*N/A*	*~3 hours*
System Requirements	*N/A*	*Nvidia GPU*

Chapter 7

Conclusion and Future Work

7.1 Summary of Findings

This thesis included mainly the Integration of Large Language Models (LLMs) into the POLAR framework pipeline, doing so to replace and replicate the NLP techniques used for extracting entities and topics and assigning attitudes from news corpus on polarizing topics. These two models were evaluated in comparison to each other: the Standard POLAR framework and the altered GPT-3.5-turbo integrated POLAR framework. Additionally, a fine-tuned Mistral-7B model trained on the GPT's outputs was evaluated on its ability to replicate it with great success.

Analyzing the evaluation through the use of qualitative and quantitative measures, it was shown that GPT in general identifies fewer overall dipole attitudes, but does so with a much stronger focus on sociopolitical topics that are typically associated with public polarization, as shown by a contextual comparison of the dipole topics. In contrast, POLAR detects a greater number of dipoles and topics, but many are shallow or of little to none sociopolitical relevance.

As for the evaluation of the Mistral model, it has successfully replicated GPT's outputs remarkably, with over 93% coverage of GPT's Entity and Topical pairs as well as roughly 97% accuracy in terms of attitude classification. In terms of Precision it was quite high at 86-87%, signaling that

when Mistral detects a pair, it is most likely a part of GPT's output, with some being Mistral exclusive pairs, which could be interpreted as enriching or introducing noise depending on the use.

As such, these findings strongly support the viability of LLMs in replacing traditional NLP components for polarization analysis and suggest that local fine-tuned models like Mistral can offer high performance while maintaining deployment flexibility, and can perhaps even outperform the standard LLMs with proper training.

7.2 Model Trade-offs

As shown by the results, GPT is set as the benchmark for LLMs ability to detect these Polarizing relationships, showcasing this in its ability to detect sociopolitical topics in the dipoles, however it does have some limitations, such as its reliance on internet access, reliance on OpenAI API availability, the usage costs and its lack of customization, able to modify only a few parameters such as Temperature.

However, Mistral has some limitations as well, such as its greater inference time per article (1-2 minutes vs GPT's 15-35 seconds) and requires an upfront effort for fine-tuning, as well as a specific GPU (Nvidia) to be able to utilize it.

As such, the choice between these systems depends on user needs and capabilities. For example, for scalable, fast and easier to setup with on-demand inference, GPT remains ideal. For a private, cost-effective and customizable option, Mistral might be better.

7.3 Limitations

This study had several constraints. The evaluation was conducted on a limited set of articles, which restricts the generalizability of the findings. The notion of what constitutes a "correct" or "meaningful" attitude or topic remains somewhat subjective, especially in politically complex or ambiguous articles. Additionally, this work only considered binary dipoles, where Entities felt

Positive/Neutral/Negative, whereas real-world discourse may involve more complex, multi-party relationships. The scope was also restricted to English-language datasets.

7.4 Future Work

- **Dynamic topic scoring**: Integrating a non-binary scale for attributing to Sentiments, as well as more expressive Attitude Dipole SAGs.
- **Dataset expansion**: Applying the models across larger datasets, containing lengthier Articles, as well as a bigger number of Articles to parse.
- **Domain Expansion**: Evaluating model outputs across different domains (healthcare, climate change, foreign policy etc.)
- **Additional Mistral Training and Evaluation**: The Fine-Tuned Mistral Evaluation was only conducted comparing itself to GPT3.5, with remarkable results. Additionally a more expansive dataset should be employed and evaluated that is not limited to GPT's outputs.

References

[1] Paschalides, Demetris, George Pallis, and Marios D. Dikaiakos. "Polar: a holistic framework for the modelling of polarization and identification of polarizing topics in news media." *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2021.

[2] Roscini, Flavia. "How The American Media Landscape is Polarizing the Country." *Frederick S. Pardee School of Global Studies* (2024).

[3] Bjornsgaard, Kelsey, and Simeon Dukić. "The Media and Polarisation in Europe: Strategies for Local Practitioners to Address Problematic Reporting." *Luxembourg: Publications Office of the European Union. Available online: https://home-affairs. ec. europa. eu/system/files/2023-05/ran_the_media_and_polarisation_052023_en. pdf (accessed on 1 December 2023)* (2023).

[4] Deng, Lingjia, and Janyce Wiebe. "Mpqa 3.0: An entity/event-level sentiment corpus." *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2015.

[5] Yarchi, Moran, Christian Baden, and Neta Kligler-Vilenchik. "Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media." Political Communication 38.1-2 (2021): 98-139.

[6] Xing, Yunfei, et al. "Diving into the divide: a systematic review of cognitive bias-based polarization on social media." Journal of Enterprise Information Management 37.1 (2024): 259-287.

[7] Bliuc, Ana-Maria, Ayoub Bouguettaya, and Kallam D. Felise. "Online intergroup polarization across political fault lines: An integrative review." Frontiers in Psychology 12 (2021): 641215.

[8] Honnibal, Matthew, et al. "spaCy: Industrial-strength natural language processing in python." (2020).

[9] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[10] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

[11] Mavroudis, Vasilios. "LangChain." (2024).

[12] OpenAI. (2023). GPT-3.5 & GPT-4 API documentation.

[13] Tripti R, Kulkarni, et al. "Deepseek Open-Source AI." International Journal of Trend in Scientific Research and Development 9.3 (2025): 555-560.

[14] Zheng, Yaowei, et al. "Llamafactory: Unified efficient fine-tuning of 100+ language models." arXiv preprint arXiv:2403.13372 (2024).

[15] Lehmann, Jens, et al. "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia." Semantic web 6.2 (2015): 167-195.

[16] Triana, Brian P., et al. "Proof-of-Concept Prompted Large Language Model for Radiology Procedure Request Routing." Journal of Vascular and Interventional Radiology (2025).

[17] Hendrycks, Dan, et al. "Measuring massive multitask language understanding." arXiv preprint arXiv:2009.03300 (2020).