Disinformation Demasked – Tech Report

Table of Contents

Disinfo demasked
Supported by the
Federal Ministry for Economic Cooperation and Development

Introduction

In this tech report, we delve into the process of developing a machine learning model. Our work encompasses data collection, data selection and preparation, machine learning experiments, and evaluation. We provide detailed steps to recreate our work, along with a list of relevant links and tools. Additionally, we highlight any limitations encountered during the process where applicable.

1. Data Collection

Description & Output

The aim of this step is to decide on a source and method for collecting the required data. The output is a tabular file containing all the data points used for labelling and LLM instruction.

Decisions made

It was decided to use the official channels of the two largest Russian news outlets, Sputnik and Russia Today, as sources of data. The timeframe was set to begin just before the start of the war in Ukraine and end at the time of collection.

- Platform:
  - Telegram
- Timeframe:
  - Start Date: 17.05.2018
  - End Date: 31.11.2023
- Sources:
  - Sputnik
    - Link: https://t.me/Sputnik_Arabic
    - Resulting number of records: ~170.000
  - Russia Today
    - Link: https://t.me/rtarabictelegram
    - Resulting number of records: ~270.000

Links & Tools

The data collection was unfortunately done with a proprietary tool of the technology partner. It is therefore not possible to share the source code. However, the steps to reproduce the work provide general instructions on how to replicate the functionality.

Steps to recreate work

Disinfo demasked
Supported by the
Federal Ministry for Economic Cooperation and Development

The data was collected using the official Telegram API. As the original source code is not publicly available, we want to at least provide instructions on how to easily replicate the data collection step so that the method can be applied to other datasets.

1. To write your own API crawler, all you need to do is follow the steps in the Telegram API documentation.
2. There are also a number of open source applications that can do this, for example on Github. These nonetheless require authentication details such as a Telegram API ID & Hash Code.
3. It is probably also possible to use an LLM to write a basic crawling script. This would also require registration for authentication details such as API ID and hash code.

2. Data Selection & Preparation

Description & Output

This section covers all processing steps performed on the database, including filtering and translation of contributions. This technical report only covers the preparation of the data up to the point of human labelling, and only mentions the labelling output. The labelling itself is part of the methodology and is therefore omitted.

Decisions made

The general processing steps include selecting a subset of the data and translating the Arabic telegram entries into English. Finally, the columns related to labelling were added and the items to be labelled by the research team were set aside.

- Data was selected based on the highest number of views and forwards.
  - The top 1000 posts in terms of both views and forwards in each channel were selected.
  - Result: 4000 unique records, Date (17.05.2018 - 31.11.2023)
- The Arabic text has been translated into English to enable a wider audience to access the data.
- To prepare the dataset for (human and machine) labelling, new columns were added to the dataframe.
  - The new columns are: Main theme, Subtheme, Confidence level, Multiple themes
  - See the research methodology for a more detailed explanation.
- The posts for human labelling were selected from the dataset of 4000 posts based on top views and forwards.
  - The 500 posts with the most engagements were selected and divided amongst the researchers.
  - Human labelling resulted in 270 entries with the four columns mentioned above.

Links & Tools

All steps are performed within Google Sheets, so no external software is required.

Disinfo demasked

Supported by the Federal Ministry for Economic Cooperation and Development

- Translation can be done within Google Sheets using the free Google Translate.
    - The English version of the text was translated using Google Translate API through Google Sheets.

Steps to recreate work

This step involves only basic table editing methods and therefore requires no further explanation.

3. Machine Learning Experiments

Description & Output

In this report, we present our initial experiments with using GPT (Generative Pre-trained Transformer) for automated data labelling. Our aim was to explore the feasibility of using GPT's capabilities to automatically assign labels to data instances. Specifically, we focused on multi-label classification cases, attempting to replicate human labelling by researchers via LLM (including confidence score and comments).
The output of this step is:

4. A recreation of the human labelled dataset by the LLM for machine evaluation.
5. Labelling of additional unseen data points for human evaluation.

Due to time constraints, this method is only a first draft and by no means optimised. Future work can focus on improving these results.

Decisions made

The general aim was to use GPT to replicate the labelling process of human researchers. While other methods such as fine-tuning or testing other (open source) LLMs were considered, due to time constraints the final decision was to perform few-shot prompting using OpenAI's GPT API.

- **Few-Shot Prompting Approach**: Our first attempt was to use few-shot prompting. This approach is known for its efficiency and quick set-up. We designed sample prompts and observed GPT's performance in labelling data based on these prompts. On top of the example codings, the prompt contained the coding instructions explaining the meaning of each theme and sub-theme.
- **Data Split and Cross-Validation**: We split our dataset of approximately 270 human annotated examples into five chunks. For each split, we labelled 20% of the data in blocks of 10 and used 15 examples from the remaining 80% to illustrate the coding methodology.
- **Labelling of Unseen Data:** Using the same approach, approximately 1500 new and unseen posts were labelled in blocks of 10, using 15 examples from the human annotated data to illustrate the coding methodology. The 1500 posts were randomly selected from the database of 4000.
- **Model used:** The GPT-4 API was used for the fivefold cross-validation. For the re-labelled data, GPT-4-turbo was used (as the API cost is lower for a presumed better model).

Disinfo demasked
Supported by the
Federal Ministry for Economic Cooperation and Development

- **Use of only the English text:** For simplicity, the database used for training only contained the English translation of the text, and the empty columns for labelling. All other fields were removed.

Links & Tools

All files, prompts and source code are available as open source. A general understanding of the OpenAI API is an advantage.

- Details about the OpenAI API can be found in the official documentation.
- All files to recreate the experiments can be found here.

Steps to recreate work

To recreate these work steps, follow the description of the steps in the code repository.
The order of executing the functions is:

6. *create_splits*: Create five 80-20 splits of the human labelled data for cross validation.
7. *call_gpt*: Loop through all data points and generate predictions via GPT. A boolean value decides whether the previously created splits or the unseen dataset should be used for labelling.
8. *Make_prediction_xlsx*: The output of the GPT call is a list of predictions in a python dictionary format, which is turned back into the original XLSX format via this function.

Evaluation

Description & Output

This step involves an automated machine evaluation of the approach and a second evaluation of the newly labelled data points by the research team. Code for the machine scoring can be found in the provided repository.
The output is numerical scores generated by a script and a database containing details of the human judgement.

Decisions made

The technical evaluation was carried out by comparing the machine-labelled data with the true human labels. The human evaluation took a more qualitative approach, focusing on providing detailed comments on the quality of the machine predictions.

- For the machine evaluation, it was decided to generate four standard evaluation metrics comparing the real and artificial labels.
  - The selected scores are: Accuracy, Precision, Recall, F1 Score
- The human evaluation was performed in the form of a random spot check.
  - Each researcher was randomly assigned 100 machine-labelled data points.
  - The database was extended to include two additional fields: Researcher comment & Final decision.

Disinfo demasked
Supported by the Federal Ministry for Economic Cooperation and Development

- The final decision was made using a multi-label scoring system (correct, partially correct, partially incorrect, incorrect).
- "Partially correct" indicates that some field was missing, while "partially incorrect" means that some theme was forgotten.
- The researcher's comment is detailed enough to recreate the correct label. However, this has not yet been done
-

Links & Tools

The files and source code to recreate the machine evaluation are available as open source. Please find them on the website.

Steps to recreate work

Details about the script can be found in the repository. To run the evaluation, use the following function:

9. *evaluate_splits*: Compares the true and predicted themes per post and calculates the scores across all data points.

Results and Comments

10. ***Machine Five-Fold Cross Validation***

Results:

- **Themes**:
  - Accuracy: 0.5290
  - Precision: 0.6538
  - Recall: 0.7349
  - F1 Score: 0.6920
- **Subthemes**:
  - Accuracy: 0.4238
  - Precision: 0.5799
  - Recall: 0.6114
  - F1 Score: 0.5953

Insights:

- **Subthemes not contained in examples are still used:** To label the unseen data, only 15 example codings were added to the prompt. As the prompt also included explanations of all themes and sub-themes, GPT was able to assign themes that were not included in the examples.
- **Subthemes not contained in humanly labelled data are still used:** There are some examples where GPT assigned sub-themes that did not appear in the humanly labelled data at all (e.g. sub-theme 4.3).

Limitations:

Disinfo demasked
Supported by the
Federal Ministry for Economic Cooperation and Development

- **Missing explanations of fields to be labelled:** GPT was given instructions on the topics, but not on the labelling method. This can make it difficult to complete the task correctly, especially for ambiguous fields such as 'confidence score'.
- **Random selection of data points as few-shot examples:** The selection of examples to instruct GPT on how to perform the task was randomised. This could introduce bias into the process, especially due to the lack of explanations (e.g. 13/15 examples with a "high" confidence score).
- **Missing time stamps in GPT prompts:** The prompt does not include the date of the post. This is important information because the time window is needed to link statements to world events at the time (e.g. sanctions to a war going on at the time).

11. ***Human Evaluation***

Results:

- Number of assigned labels per category for each post:
  - Correct: 208
  - Partially correct: 57
  - Partially incorrect: 5
  - Incorrect: 14
- Ratio of Labels:
  - 73% were correct, 93% were at least partially correct
  - 5% were incorrect, 7% were at least partially incorrect

Insights:

- **Overall positive results:** GPT managed to assign the correct, or at least partially correct labels in the majority of validated samples.
- **Limitations noticeable in qualitative analysis:** Some of the errors GPT made could be explained via the aforementioned limitations of the approach, such as the missing time-stamps (e.g. leading to sanctions against Russia not being tied to war in Ukraine happening at the time).

Limitations (on top of the above):

- **Lack of numeric evaluation scores:** True labels were not created, so it is not possible to calculate evaluation metrics. However, this can be done in the future as all the necessary information is contained in the researcher's comments.

Disinfo demasked
Supported by the
Federal Ministry for Economic Cooperation and Development